Anja Nedoluzhko

Discourse

Annotation

Theories and Tools

Corpora

Anja Nedoluzhko
What today’s lecture is NOT about:

0 computational processing of discourse, discourse parsing of any type
0 event annotation (Martha Palmer, James Pustejovsky et al.)
0 various forms of connectives (DRDs) in different languages (other classes)
0 just a little about spoken corpora
What today’s lecture IS about

I. Very generally about discourse-related topics
II. Discourse theories (H&H, RST, SDRT, PDTB, DGB, CCR) and corpora annotated with discourse relations

Overview of discourse theories will be made from the point of view of relations and conceptions (not connectives, they will be addressed within other courses)

III. Comparison of theories and approaches
I. General words about discourse-related topics

- large/small discourse units,
- coreference,
- cohesion/coherence,
- spoken/written,
- elementary discourse units,
- requirements to discourse structure
• Logical document structure

*a hierarchy of segments of the document, each of which corresponds to a visually distinguished semantic component of the document (Summers 1998)*
Processing Large Discourse Units

- Logical document structure
  
a hierarchy of segments of the document, each of which corresponds to a visually distinguished semantic component of the document (Summers 1998)
Processing Large Discourse Units

- Logical document structure
  a hierarchy of segments of the document, each of which corresponds to a visually distinguished semantic component of the document (Summers 1998)

- Content zones
  a continuous portion of a text document that fulfills a functional role for the text as a whole, contributing to the overall message or purpose, as it is characteristic for the genre of the text (defined for each sentence)

Teufel and Moens (2002) - aimed at recovering content zones from conference papers in CL.
- Aim, research goal
- Textual: statements about section structure
- Own: description of the authors' work (methodology, results, discussion)
- Background: generally accepted scientific background
- Contrast: comparison with other work
- Basis: statements of agreement with other work
- Other: description of other researchers' work

Synthesis of pyrazole and pyrimidine Troeger's base-analogues

Rodrigo Abeniga, Andrés Alberola, Hector Larranaga, Mário Quirino, Brazio Inassaty, Henry Inassaty, Angelina Hawnam, Adolfo Benitez, and Manuel Noguera

Troeger's-base analogues bearing fused pyrazole or pyrimidine rings were prepared in acceptable to good yields through the reaction of 3-alkyl-5-aminopyrazoles or 4-aminoimidazoles with maleimide under solid conditions (i.e. in ethanol at SC, in the presence of catalytic amounts of acetic acid). Two key intermediates were isolated from the reaction mixture, which helped us to suggest a sequence of steps for the formation of the Troeger's bases obtained. The structures of the products were assigned by 1H and 13C NMR, mass spectra and elemental analysis and confirmed by X-ray diffraction for one of the obtained compounds.

Introduction

Although the first Troeger's base \cite{1} was obtained more than a century ago from the reaction of p-toluidine and formaldehyde \cite{1}, recently the study of these compounds has gained importance due to their potential pharmaceutical applications. They possess a relatively rigid chiral structure which makes them suitable for the development of possible synthetic enzymes and artificial receptor systems \cite{2}, chelating and bimetallic systems \cite{3} and transition metal complexes for angled acid-precipitable catalystic reactions \cite{4}. For these reasons, numerous Troeger's-base definitions have been prepared bearing different types of substituents and structures (i.e. 2,5- or 3,6-Substitution) with the purpose of increasing their potential applications \cite{5,6}.

Scheme 1. The original Troeger's base \(1\) and some interesting derivatives and analogues.

However, some of the above methodologies possess tedious work-up procedures or include relatively strong reaction conditions, such as treatment of the starting materials for several hours with an alcoholic solution of conc., hydrochloric acid or TFA solution, with poor to moderate yields, or in the case for analogues \(4\) and \(5\).

Considering these potential applications, we now report a simple synthetic method for the preparation of \(5,12\)-dicyclo-\(3,10\)-dicycloc-3,6,13,16(tetrahydrobenzo[b]thiepin-2(3\(H\)))-1,4,11,13-tetramines \(9\) and \(4,12\)-dimethoxy-3,5,9,11,13,15-hexahydrobenzo[b]thiepin-2(3\(H\)))-1,4,11,13-tetramines \(10\) and \(11\) obtained with maleimide in ethanol and catalytic amounts of acetic acid. Compounds \(8\) and \(10\) are new Troeger's base analogues bearing heterocyclic rings instead of the usual phenyl rings in their aromatic parts.

Results and discussion

In an attempt to prepare the benzotroenyl-derivatives \(7\), which could be used as an intermediate in the synthesis of new \(1\), such as a mixture of \(5\)-amino-3-methyl-1-phenylpyrazole \(1\), formaldehyde and benzylamine, in \(30\) ml of ethanol, with catalytic amounts of acetic acid, was treated at SC for \(5\) minutes. A solid precipitated from the solution which was still hot. However, no consumption of benzylamine was observed at TLC.

The reaction conditions were modified and the same product was obtained when the reaction was carried out without using benzylamine, as shown in Scheme 12. On the basis of NMR and mass spectra and X-ray crystallographic analysis we established that the structure is \(3,12\)-dicyclo-\(3,10\)-dicyclo-3,6,13,16(tetrahydrobenzo[b]thiepin-2(3\(H\)))-1,4,11,13-tetramine \(13\) or demethylated \(5\). In addition, \(8\), a new pentacyclic Troeger's base analogue.

© Manfred Stede / NAACL Tutorial 2013
Processing Large Discourse Units

- **Logical document structure**
  
a hierarchy of segments of the document, each of which corresponds to a visually distinguished semantic component of the document (Summers 1998)

- **Content zones**
  
a continuous portion of a text document that fulfills a functional role for the text as a whole, contributing to the overall message or purpose, as it is characteristic for the genre of the text (defined for each sentence)

- **Topic-based segmentation**
  
a sequence of non-overlapping text segments that completely covers the text. Each unit consists of one or more sentences that address a common topic.

---

**Discourse Processing**

[1.7] A man named Lionel Gaedi went to the Port-au-Prince morgue in search of his brother, Josef, but was unable to find his body among the piles of corpses that had been left there. [1.8] “I don’t see him—it’s a catastrophe,” Gaedi said. [1.9] “God gives, God takes.” [1.10] Chris Rolling, an American missionary and aid worker, tried to extricate a girl named Jacqueline from a collapsed school using nothing more than a hammer. [1.11] He urged her to be calm and pray, and as night fell he promised that he would return with help. [1.12] When he came back the next morning, Jacqueline was dead. [1.13] “The bodies stopped bothering me after a while, but I think what I will always carry with me is the conversation I had with Jacqueline before I left her,” Rolling wrote afterward on his blog. [1.14] “How could I leave someone who was dying, trapped in a building! ...[1.15] She seemed so brave when I left! [1.16] I told her I was going to get help, but I didn’t tell her I would be gone until morning. [1.17] I think this is going to trouble me for a long time.” [1.18] Dozens of readers wrote to comfort Rolling with the view that his story was evidence of divine wisdom and mercy.

© Manfred Stede / NAACL Tutorial 2013
**Processing Large Discourse Units**

- **Logical document structure**
  
  A hierarchy of segments of the document, each of which corresponds to a visually distinguished semantic component of the document (Summers 1998)

- **Content zones**

  A continuous portion of a text document that fulfills a functional role for the text as a whole, contributing to the overall message or purpose, as it is characteristic for the genre of the text (defined for each sentence)

- **Topic-based segmentation**

  A sequence of non-overlapping text segments that completely covers the text. Each unit consists of one or more sentences that address a common topic

- **Discourse modes (C. Smith, 2003)**

  Narrative, information, argument (commentary), report, description

**Determining DMs in NLP are good for:**

- Temporal discourse processing,
- Automatic summarization,
- Information extraction,
- Argumentation mining,
- Genre distinctions, ...

**Situation entity types (A. Friedrich)**

- State
- Event
- Generic sentence
- Generalizing sentence
- General stative
- Abstract entity
- Report
- Speech act

**Narrative**

The chicken or the egg causality dilemma is commonly stated as “which came first, the chicken or the egg?”. To ancient philosophers, the question about the chicken or egg also evoked the question of how life and the universe in general began.

In my opinion, the results of Prof. Dr. Origin and his group are highly interesting. However, no by no means solve the philosophical question of life and the universe. I believe that much more research is needed in the field of biology alone will not be able to answer this question.

---

**REPORT**

**State, Event**

Temporal progression related to speech time.

On Monday, NASA announced that signs of liquid water have been found on Mars. The Mars Reconnaissance Orbiter spacecraft found evidence of the liquid on the Martian surface, in long dark spots on the Red Planet thought to be formed because of water flow.

**DESCRIPTION**

**State, on-going Event**

Metaphorical progression through scene

The sandhills here run down to the sea, and end in two spits of rock jutting out opposite each other, till you lose sight of them in the water. One is called the North Spit, and one the South.

Annemarie Friedrich. Annotation and automatic classification of situation entity types. 2015

---

**Argument**

**Commentary**

one text

\[ \approx \text{one genre} \]

one passage

\[ \approx \text{one discourse mode} \]
• Logical document structure
  a hierarchy of segments of the document, each of which corresponds to a visually distinguished semantic component of the document (Summers 1998)

• Content zones
  a continuous portion of a text document that fulfills a functional role for the text as a whole, contributing to the overall message or purpose, as it is characteristic for the genre of the text (defined for each sentence)

• Topic-based segmentation
  a sequence of non-overlapping text segments that completely covers the text. Each unit consists of one or more sentences that address a common topic

• Discourse modes (C. Smith, 2003)
  narrative, information, argument (commentary), report, description

Coreference

  grammatical coreference
    • mostly possible to identify the antecedent on the basis of grammatical rules of the given languages
    • within one sentence

  * arguments in constructions with verbs of control
  (John wants to [#Cor.ACT] kiss Mary.)

  * reflexive pronouns
  (John shaved himself)

  * relative pronouns
  (John, who came late, apologized.)

  * coreference with verbal modifications that have dual dependency
  (John saw Mary [#Cor.ACT] stand on the windowsill and cry.)

  * reciprocity
  (John and Mary kissed [#Rcp.PAT].)
• Logical document structure
  a hierarchy of segments of the document, each of which corresponds to a visually distinguished semantic component of the document (Summers 1998)

• Content zones
  a continuous portion of a text document that fulfills a functional role for the text as a whole, contributing to the overall message or purpose, as it is characteristic for the genre of the text (defined for each sentence)

• Topic-based segmentation
  a sequence of non-overlapping text segments that completely covers the text. Each unit consists of one or more sentences that address a common topic

• Discourse modes (C. Smith, 2003)
  narrative, information, argument (commentary), report, description

Coreference

  • grammatical coreference
  • textual coreference
    • not restricted to grammatical means alone, context
    • different means (pronominalisation, grammatical agreement, repetitions, synonyms, paraphrasing, hyponyms/hyperonyms, etc.)
    • often occurs between entities in different sentences

  * personal and possessive pronouns (*Jonh left Mary. *He wanted to see his mother*),
  * demonstrative pronouns *ten, ta, to (It means that he doesn’t really love Mary.)*
  * with textual ellipsis (zeros) (*Vice si Ø vážil své matky.*)
  * nouns (*John asked his mother to advise him how he should behave with Mary. but mother ignored her son’s wish.*)
  * local adverbs (*John asked mother to come to Mary’s place with him but she decided not to go there.*)
  * some adjectives (*At last, Mary came to Prague herself and found the Prague atmosphere quite casual.*)
  * reference to events (*Mary suggested Jonh to go to the theater, but Jonh ignored her wish.*)
  * If antecedent is a whole segment of (previous) text larger than one sentence (phrase) — special type of textual coreference segm(ent) without explicitly marked antecedent: (*The next day Mary suggested to visit his mother. Then she proposed to go swimming. Her last wish was just to look at the city center. Jonh denied all of it.*)
Processing Large Discourse Units

- Logical document structure
  a hierarchy of segments of the document, each of which corresponds to a visually distinguished semantic component of the document (Summers 1998)

- Content zones
  a continuous portion of a text document that fulfills a functional role for the text as a whole, contributing to the overall message or purpose, as it is characteristic for the genre of the text (defined for each sentence)

- Topic-based segmentation
  a sequence of non-overlapping text segments that completely covers the text. Each unit consists of one or more sentences that address a common topic

- Discourse modes (C. Smith, 2003)
  narrative, information, argument (commentary), report, description

Coreference resolution

Processing Small Discourse Units

- coherence relations
- connectives, cue phrases

our focus for today
In a cohesion analysis the connectivity of the discourse is primarily tied to the explicit marking of semantic relations. These explicit cues make a text a text.

BUT: Cohesion is necessary but not sufficient condition for the creation of text (Halliday and Hasan, 1976).

Cohesive elements like connectives are viewed as important but not necessary features of discourse: they are linguistic markers expressing the underlying conceptual relations that are of a cognitive nature (Sanders, 1992).

Texture (quality that makes a particular set of words and sentences a text)

Coherence and cohesive relations are the threads that make up a text.

Discourse is coherent if

- all of its pieces belong together
- and they have some common thread

coherence (relational coherence)

cohesion (entity-based coherence)
Requirements to a theory of discourse structure

A satisfying theory of discourse structure should meet:

- **Descriptive Adequacy**: A theory discourse structure makes it possible to describe the structure of all kinds of natural texts.

- **Psychological Plausibility**: A theory of discourse structure should at least generate plausible hypotheses on the role of discourse structure in the construction of the cognitive representation.

(\textit{it should make sense} 😊)

Ted Sanders (1992)
Elementary Discourse Units (EDUs)

Give me your essay, please. Then you may leave the room.

Give me your essay, please. The best one.

When you leave the room, give me your essay.

John attended the lecture despite his illness.

Exhausted, I called for a taxi cab.

John, although he was ill, attended the lecture.

The red car that's parking in front of you belongs to me.

The red car, which was brought here by my dad, belongs to me.
Elementary Discourse Unit (EDU)

A span of text, usually a clause, but in general ranging from minimally a (nominalization) NP to maximally a sentence. It denotes a single event or type of event, serving as a complete, distinct unit of information that the surrounding discourse may connect to. An EDU may be structurally embedded in another.

Abstract objects (Asher, 1993)

- events, states, conditions and dialogue acts, that are typically expressed as sentences, but they can also be smaller or larger units (clauses, paragraphs, dialogue segments)

BUT!

Vary from researcher to researcher, depending on the level of granularity needed

Pragmatic elementary units (Polanyi et al. 2004)

- meaning,
- discourse function,
- independent continuation

Interprovincial Pipe Line Co. said it will delay a proposed two-step, 830 million dollar (US$705.6 million) expansion of its system because Canada's output of crude oil is shrinking.

Elementary Discourse Unit (EDU)

- basic intuition – they also exist underneath the clausal level, and they can either be titles, oppositions, relative clauses, some NP (e.g. principles), etc.

[Interprovincial Pipe Line Co. said] \(\pi_1\)

[It will delay a proposed two-step, 830 million dollar] \(\pi_3\) expansion of its system] \(\pi_2\) [because Canada’s output of crude oil is shrinking.] \(\pi_4\)

Nicolas Asher

[Principes de la sélection naturelle.] \(\pi_1\) [La théorie de la sélection naturelle [telle qu'elle a été initialement décrite par Charles Darwin.] \(\pi_3\) repose sur trois principes :] \(\pi_2\) [1. le principe de variation] \(\pi_4\) [2. le principe d'adaptation] \(\pi_5\) [3. le principe d'hérité] \(\pi_6\)

Title, NP

Apposition
**Written vs. Spoken**

0 **Written text** - of a single author

0 The author is organizing his thoughts together, make them coherent as much as he can

---

0 **Conversation** as a game of message exchange involving a kind of signaling game, a play with reactions:

  0 X plays $\varphi \rightarrow$ Y decodes a message $\rightarrow$ Y decides what signal to send in return $\rightarrow$ X decodes a message.

---

Definitely different!

- a number of discourse structure annotated corpora for dialogues (Asher’s STAC, Sidarenka (chats), L. Degand’s LOCAS-F, Italian?)
- annotating discourse on spoken texts requires first of all annotation of DMs.
I. Very generally about discourse-related topics

II. Discourse theories (H&H, RST, SDRT, PDTB, DGB, CCR) and corpora annotated with discourse relations

III. Comparison of theories and approaches
THEORIES & CORPORA

- Halliday and Hasan – Cohesion in English
- RST – Rhetorical Structure Theory
- SDRT - Segmented Discourse Representation Theory
- PDTB – Penn Discourse Treebank
- Discourse Graphbank
- CCR – Cognitive approach to coherence relations
We can distinguish what is text and what is not → text has structure

But the unity of a text is a unity of a different kind

Text is realized by sentences but it does not have the same structural integration among its parts as we find among the parts of a sentence or clause

Investigate the resources that English has to create text

Cohesion – a set of explicit cues – is what makes a text a text
**Types of cohesion:**

0 **Reference**: identity between entities

*John loves Mary. However he is afraid to kiss her.*

0 **Substitution**: similarity between different instantiated entities of the same type

*These biscuits are stale. Get some fresh ones.*

0 **Ellipsis**

*Would you like to hear another verse? I know twelve Ø more.*

0 **Conjunction**: logico-semantic relations between propositions (e.g. addition, contrast, cause)

*John love Mary. However he is afraid to kiss her.*

0 **Lexical cohesion** (similarity between entities of the same type based on sense relations (e.g. hyperonymy, part-whole relations)

*Why does this little boy wriggle all the time? Girls don’t wriggle.*
German-English Contrasts in Cohesion (GECCo)

- based on the definition of cohesion and cohesive devices in English by (Halliday & Hasan, 1976)
- elaborated for a contrastive analysis of two languages
- comparable and parallel texts in English and German (ca. 80,000 sentences)
- various registers, including written and spoken dimensions

German-English Contrasts in Cohesion (GECCo)

- reference: personal head, modifier, *it/es*-endo- and -exophoric, demonstrative head, modifier, local, temporal, comparative particular and general

- conjunctive relations: additive, adversative, causal, temporal, modal

- substitution: nominal, verbal, clausal

- ellipsis: nominal, verbal, clausal

- lexical cohesion: general nouns, repetition, synonymy, antonymy, hyperonymy, hyponym, meronymy

- chains: Nr of chains, chain length, Nr of antecedents

THEORIES & CORPORAS

0 Halliday and Hasan – Cohesion in English
0 RST – Rhetorical Structure Theory
0 SDRT - Segmented Discourse Representation Theory
0 PDTB – Penn Discourse Treebank
0 Discourse Graphbank
0 CCR – Cognitive approach to coherence relations


http://www.sfu.ca/rst/
http://www.sfu.ca/rst/05bibliographies/
Rhetorical Structure Theory

- Empirical perspective, comes from corpus analysis
- RST is about how text works (primarily written → spoken)

  ✓ Coherent texts consist of minimal units, which are linked to each other, recursively, through rhetorical relations (coherence or discourse relations)

  ✓ Coherent texts do not show gaps or non-sequiturs.

  ✓ The resulting structure is a complete tree: no cross-dependencies.
    - completedness (one schema application contains the entire text)
    - connectedness (each span, except for the span that contains the entire text, is either a minimal unit or a constituent of another schema application)
    - uniqueness (each schema application contains a different set of text spans)
    - adjacency (the spans of each schema application constitute one contiguous text span)
RST: Graphical Representation

- A horizontal line covers a span of text (possibly made up of further spans)
- A vertical line signals the nucleus or nuclei
- A curve represents a relation, and the direction of the arrow, the direction of satellite towards nucleus
Relations in RST

- RST describes coherence relations between discourse segments: RST analysis wants to answer the question how coherence in text is achieved.
- Definition of the relations are based on functional and semantic criteria, not on morphological or syntactic signals.
- No reliable or unambiguous signal for any of the relations.
- Different lists of the relations exist:
  - Mann and Thomson – 24 relations (Cause, Contrast, Elaboration, Restatement, Evidence, Conditions, Antithesis...)
  - Later - List, Means, Preparation, Unconditional, Unless, two Restatements (nuclear and multinuclear), Joint (the declared absence of a relation)
<table>
<thead>
<tr>
<th>Relation names (Mann-Thompson 1988)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circumstance</td>
</tr>
<tr>
<td>Solutionhood</td>
</tr>
<tr>
<td>Elaboration</td>
</tr>
<tr>
<td>Background</td>
</tr>
<tr>
<td>Enablement and Motivation</td>
</tr>
<tr>
<td>Enablement</td>
</tr>
<tr>
<td>Motivation</td>
</tr>
<tr>
<td>Evidence and Justify</td>
</tr>
<tr>
<td>Evidence</td>
</tr>
<tr>
<td>Justify</td>
</tr>
<tr>
<td>Relations of Cause</td>
</tr>
<tr>
<td>Volitional Cause</td>
</tr>
<tr>
<td>Non-Volitional Cause</td>
</tr>
<tr>
<td>Volitional Result</td>
</tr>
<tr>
<td>Non-Volitional Result</td>
</tr>
<tr>
<td>Purpose</td>
</tr>
</tbody>
</table>

Other classifications are possible, both longer and shorter lists have been proposed.
Relations in RST - nuclearity

Coherence relations between discourse segments

• asymmetric ("mononuclear")
  • one nucleus, one satellite
• symmetric ("multinuclear")
  • multiple nuclei

other schemas
Hypotactic (subordinate)

0 Sub-sentential Concession relation

0 Concession across sentences

0 Nucleus (spans 2-3) made up of two spans in an Antithesis relation
Paratactic (coordinate)

• At the sub-sentential level (traditional coordinated clauses)
  Peel oranges, and slice crosswise.

• But also across sentences
  1. Peel oranges, 2. and slice crosswise. 3. Arrange in a bowl 4. and sprinkle with rum and coconut. 5. Chill until ready to serve.
A relation consists of:
1. Constraints on the Nucleus,
2. Constraints on the Satellite,
3. Constraints on the combination of Nucleus and Satellite,
4. The Effect.

1. Constraints on the Nucleus
   The reader may not believe N to a degree satisfactory to the writer
2. Constraints on the Satellite
   The reader believes S or will find it credible
3. Constraints on the combination of N+S
   The reader’s comprehending S increases their belief of N
4. Effect (the intention of the writer)
   The reader’s belief of N is increased

Example: Evidence

1) Darwin as a Geologist 2) He tends to be viewed now as a biologist, 3) but in his five years on the Beagle his main work was geology, 4) and he saw himself as a geologist. 5) His work contributed significantly to the field.
RST corpora

0 RST Discourse Treebank (Carlson et al. 2003)
0 RST Signalling Corpus (Das and Taboada 2015)
0 Potsdam Commentary Corpus (Stede and Neumann 2014)
  0 220 German newspaper commentaries annotated with different types of linguistic information, including RST
0 Discourse Relations Reference Corpus
  0 http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html
  0 texts from RST web site + annotated Wall Street Journal articles from the RST Discourse Treebank + review texts from the SFU Review Corpus
0 GUM - The Georgetown University Multilayer Corpus
  0 POS, lemmas, syntax, constituent and dependency syntax, Information status (given, accessible, new)
  0 Entity and coreference annotation
0 Spanish RST Discourse Treebank (da Cunha, Iria, Juan Manuel Torres-Moreno and Gerardo Sierra 2011))
  0 http://corpus.iingen.unam.mx/rst/
0 Basque RST Discourse Treebank (Iruskieta et al.)
  0 http://ixa2.si.ehu.es/diskurtsoa/en/
0 Multiling RST Treebank (English, Spanish and Basque)
  0 http://ixa2.si.ehu.es/rst
RST Discourse Treebank


- 385 WSJ articles from Penn Treebank, representing over 176,000 words of text. ~14% were double-tagged.
- Document length: 31 to 2124 words; average of 458.14 words
- Average # EDUs per document: 56.59.
- Average # words per EDU: 8.1.
- Nature of articles: general news, financial, business, cultural reviews, editorials
- Intended users: developers of automatic text processing systems
[Still, analysts don’t expect the buy-back to significantly affect per-share earnings in the short term.][16] [“The impact won’t be that great,”][17] said Graeme Lidgerwood of First Boston Corp. [This is in part because of the effect][19] [of having to average the number of shares outstanding,][20] [she said.][21] [In addition,][22] [Mrs. Lidgerwood said,][23] [Norfolk is likely to draw down its cash initially][24] [to finance the purchases][25] [and thus forfeit some interest income.][26]
Hypothesis: There are no (or very few) ‘implicit’ relations.
RST - Present and Future

Annotated corpora in different languages, maintained web

**RST workshops:**

- 2007 - Collocated with the *16th Intercâmbio de Pesquisas em Linguística Aplicada (InPLA)*, Pontifícia Universidade Católica de São Paulo, Brazil.
- 2009 - Named "Brazilian RST Meeting", collocated with *The 7th Brazilian Symposium in Information and Human Language Technology (STIL)*, in São Carlos, Brazil.
- 2011 - 3rd Workshop, "RST and Discourse Studies" was held with *The 8th Brazilian Symposium in Information and Human Language Technology (STIL)*, in Cuiabá, Brazil.
- 2013 - The 4th workshop "RST and Discourse Studies" was also held within *STIL (2013)* in Fortaleza, Brazil.
- 2015 - The 5th workshop "RST and Discourse Studies", collocated with *The 31st Conference of the Spanish Association for Natural Language Processing*. Alicante, Spain.
  - [https://sites.google.com/site/workshoprst2015/](https://sites.google.com/site/workshoprst2015/)
RST done!
THEORIES & CORPORA

- Halliday and Hasan – Cohesion in English
- RST – Rhetorical Structure Theory
- Discourse Graphbank
- SDRT - Segmented Discourse Representation Theory
- PDTB – Penn Discourse Treebank
- CCR – Cognitive approach to coherence relations
Jerry R. Hobbs’ theory


Coherence relations, e.g. *contrast, elaboration, parallel*...

- ... hold between segments of a discourse
- ... are defined in terms of propositions that can be inferred from the assertions of discourse segments (in terms of formal logic)
  - The assertion of a clausal discourse segment is, roughly, what is predicated by the main verb
Jerry R. Hobbs’ theory

**Parallel:**
Infer $p(a_1, a_2, \ldots)$ from the assertion of $S_0$ and $p(b_1, b_2, \ldots)$ from the assertion of $S_1$, where $a_i$ and $b_i$ are similar, for all $i$.

\[
\begin{array}{c|c|c}
&S_0 & S_1 \\
p & a_1 & a_2 \\
\end{array}
\]

**Elaboration:** Infer the same proposition $P$ from the assertions of $S_0$ and $S_1$.

*Go down First Street. Just follow First Street three blocks to A Street.*

go(Agent: you, Goal: x, Path: First St., Measure: y) for some x and y.
go (Agent: you, Goal: A St., Path: First St., Measure: 3 blks)

If we assume that $x$ is A Street and $y$ is 3 blocks, then the two are identical and serve as the proposition $P$ in the definition.
Discourse Graphbank


0 motivated by Hobbs (1985) and Kehler (2002)
0 vs. RST: polemics to tree structures
0 135 texts annotated with coherence relations (WSJ and AP Newswire)
0 DUs – mainly clauses
  0 delimiting discourse segments (*because, and, for example*, periods, semicolons, commas, etc.)
  0 treat attributions (*John said that …*) as discourse segments
0 Discourse Segment Groupings
  0 groups: e.g. attributed to the same source, topically related, might also be subgroups, consisting of several discourse segments each (making a partially hierarchical structure for the text)
  0 groups were allowed to partially overlapped, but this was not used by annotators
Hobbs’ relations:
- Occasion Relation
- Cause
- Evaluation Relation
- Ground-Figure
- Explanation Relations
- Expansion Relations
- Background
- Parallel
- Elaboration
- Exemplification:
  - Contrast
  - Violated Expectation

**Relations**

- Temporal sequence
- Cause – Effect
- Condition
- Elaboration
- Example
- Generalization
- Similarity
- Contrast
- Violated Expectation
- Attribution
- Same

**Asymmetrical (directed)**

following (Mann and Thompson 1988; Marcu 2000)

**Symmetrical (undirected)**
0. Farm prices in October edged up 0.7% from September
1. as raw milk prices continued their rise,
2. the Agriculture Department said.
3. Milk sold to the nation’s dairy plants and dealers averaged $14.50 for each hundred pounds,
4. up 50 cents from September and up $1.50 from October 1988,
5. the department said.
DGB vs RST (or graphs vs. trees)

Markus Egg and Gisela Redeker, How Complex is Discourse Structure? LREC 2010

DGB:
- crossed dependencies
- 41.22% of the segments have multiple parents (W&G 2005)

Egg and Redeker:
distinguish the complexity inherent in the data and the one arising from specific design choices in W&G’s annotation (on 14 texts from DGB)

Result: Many non-tree-like dependencies may be converted to tree-like ones.

(4) \((C_1)\) “He was a very aggressive firefighter. \((C_2)\) He loved the work he was in,” \((C_3)\) said acting Fire Chief Larry Garcia. \((C_4)\) “He couldn’t be bested in terms of his willingness and his ability to do something to help you survive.” (ap-890101-0003)
Discourse Graphbank done!
THEORIES & CORPORA

0 Halliday and Hasan – Cohesion in English
0 RST – Rhetorical Structure Theory
0 Discourse Graphbank
0 PDTB – Penn Discourse Treebank
0 SDRT - Segmented Discourse Representation Theory
0 CCR – Cognitive approach to coherence relations
Penn Discourse Treebank
http://www.seas.upenn.edu/~pdtb/

0 Authors, background, goals
0 Versions
0 Annotation principles
0 Related corpora

performed by Lucie Poláková
PDTB – Authors

Prof. Aravind Joshi
University of Pennsylvania

Dr. Rashmi Prasad
University of Wisconsin-Milwaukee
(formerly UPenn)

Prof. Bonnie Webber
University of Edinburgh

UPenn Collaborators (among others):

Alan Lee
Eleni Miltsakaki
Nikhil Dinesh
Premise: The meaning and coherence of a discourse results partly from how its constituents relate to each other.

DISCOURSE RELATIONS: Semantic, “informational”, relations between abstract objects (AOs) mentioned in discourse.

Abstract objects – events, states, propositions (Asher 1993)
PDTB – Background

0 **Low-level:** “Shallow” or “local” discourse analysis: Lack of agreement on high-level discourse representation structures (trees, graphs...)

0 **Theory-neutral:** Allows corpus to be usable with different frameworks; allows for data-driven “emergent” theory of discourse structure.

0 **Lexically grounded:** relations anchored by lexical items

0 **Stand-off** representation: can be easily merged with other annotations

performed by Lucie Poláková
PDTB – Goals

To annotate a large-scale corpus of discourse relations to:

0 Extend the scope of discourse-level NLP research and resulting applications;

0 Facilitate cross-linguistic empirical research on discourse relations.

performed by Lucie Poláková
Penn Discourse Treebank - History

0 PDTB 1.0 – 2004 (Miltsakaki et al., 2004, LREC Portugal)
0 PDTB 2.0 – 2008 (Prasad et al., 2008, LREC Morocco)
0 PDTB 3.0 – work in progress, two NSF grants 2014-2017 (UPenn, University Wisconsin-Milwaukee)

performed by Lucie Poláková
Annotation principles

What is annotated in PDTB?

0 Discourse relations – introduced by:
  0 Discourse connectives “but, then, for example, although…”
  0 Alternative lexicalizations of the connectives “the reason is”
  0 Implicit relations (no connective present)

0 Senses (semantics) of the relations – contrast, condition, cause...

0 Attribution “he said, I believe…”

performed by Lucie Poláková
A discourse connective is a discourse-level predicate taking two (and only two) text units (abstract objects) as its arguments:

Explicit connectives:

She hasn’t played any music since the earthquake hit.

performed by Lucie Poláková
Implicit relations:

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

AltLex:

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.
Senses (PDTB 2.0)

- Three-level hierarchy:
  - 4 classes, 15 types, 23 subtypes

Substantially revised for the 3.0 version!!

performed by Lucie Poláková
Attribution

= ascription of texts contents to agents that expressed them
0 NOT CONSIDERED A DISCOURSE RELATION in the PDTB
0 annotated for every relation and for each of its two arguments
0 4 features of attribution recognized:

Source – writer, other person, arbitrary, inherited (for arguments)

Type – verbs of communication (say, claim, explain...), propositional attitude (think, suppose...), factive (regret, remember...), control verbs (persuade, promise, intend...)

Scopal polarity – the attribution verb reverses the polarity of the argument (deny, not think...)

Determinacy – indeterminate if the attribution can be cancelled (it cannot be said...)

performed by Lucie Poláková
What is newly annotated in PDTB 3.0?

- **Intra-sentential relations!**
  - free adjuncts (-ing + ed forms) with no explicit connectives
    
    *Exhausted, I called for a taxi cab.*
    
    = *As a result of being exhausted...*
  - VP coordinations
  - Subordinators (also prepositions): “by, despite, because of, instead of, in order to, to...”
  - Punctuation marks: colon, semicolon, dash
  - Implicit relations across paragraph boundaries
  - Extending some connectives: then – but then

*performed by Lucie Poláková*
New sense hierarchy

0 Only **two levels** (subtype level cancelled, some fine distinctions not made anymore)

0 Level three only preserved via **directionality** of the relations

0 Pragmatic domain: **epistemic and speech act** readings

0 **New senses** introduced: purpose, similarity, negative condition...

0 Some senses redefined (conjunction etc.)

performed by Lucie Poláková
Corpora inspired by the PDTB annotation

- The BioDiscourse Relation Bank (BioDRB, Prasad et al., 2011) – English
- Hindi Discourse Relation Bank (HDRB, Kolachina et al., 2012, Oza et al. 2009)
- The Leeds Arabic Discourse Treebank (Al-Saif and Markert, 2010)
- PDTB-style annotation of Chinese (Zhou and Xue, 2012)
- Turkish Discourse Bank (Zeyrek et al., 2010)
- Prague Dependency Treebank 3.0 (Bejček et al., 2013) – Czech
- LUNA: PDTB-style annotation of Italian spoken dialogs (Tonelli et al., 2010)
- Potsdam Commentary Corpus (Stede 2004, Stede and Neumann 2014) – German
- French Discourse Treebank (Danlos et al. 2012)
- Tüba-D/Z Treebank (Gastel et al 2011, Versley and Gastel 2012) – German
  (specific connectives, partly implicit relations)

performed by Lucie Poláková
Literature

Most important:

0 The PDTB research group: *The PDTB 2.0 Annotation manual*, 2007.
0 Prasad et al. 2008: *The Penn Discourse Treebank 2.0*, LREC, Morocco.

All related literature available at:
http://www.seas.upenn.edu/~pdtb/bibliography-year.shtml

performed by Lucie Poláková
PDTB done!
THEORIES & CORPORÆ

0 Halliday and Hasan – Cohesion in English
0 RST – Rhetorical Structure Theor
0 Discourse Graphbank
0 PDTB – Penn Discourse Treebank
0 SDRT - Segmented Discourse Representation Theory
0 CCR – Cognitive approach to coherence relations
Segmented Discourse Representation Theory (SDRT)

Primarily based on Discourse Representation Theory (DRT: Kamp, 1981; Kamp & Reyle, 1993)

Comes from formal semantics, initially from formal sentence semantic and then systematically extended to discourse

interested in ‘semantic scopes’ – some sort of bijection between one structure to another between the relations

Recursive Structure (SDRS) - constituted by discourse units linked by discourse relations where a discourse unit is either an elementary discourse unit (EDU) or a complex discourse unit (CDU)

DUs can be embedded one into another (EDUs)

CDUs may not partially overlap


Segmented Discourse Representation Theory (SDRT)

A discourse structure in SDRT:
- $A$ is a set of labels (units)
- $\text{Last}$ is a label $A$ (intuitively last clause)
- $F$ is a formula which assigns each member of $A$ a member of a formula of the SDRS language

$$\text{SDRS} = \langle A, F, \text{Last} \rangle$$

\[ \pi_1 \text{ John had a great evening last night.} \]
\[ \pi_2 \text{ He had a great meal.} \]
\[ \pi_3 \text{ He ate salmon.} \]
\[ \pi_4 \text{ He devoured lots of cheese.} \]
\[ \pi_5 \text{ He then won a dancing competition.} \]

\[ \langle A, F, \text{LAST} \rangle, \text{ where:} \]

- $A = \{\pi_0, \pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7\}$
- $F(\pi_1) = K_{\pi_1}, F(\pi_2) = K_{\pi_2}, F(\pi_3) = K_{\pi_3}$,
- $F(\pi_4) = K_{\pi_4}, F(\pi_5) = K_{\pi_5}$,
- $F(\pi_0) = \text{Elaboration}(\pi_1, \pi_6)$,
- $F(\pi_6) = \text{Narration}(\pi_2, \pi_5) \land \text{Elaboration}(\pi_2, \pi_7)$,
- $F(\pi_7) = \text{Narration}(\pi_3, \pi_4)$
- $\text{LAST} = \pi_5$
Segmented Discourse Representation Theory (SDRT)

0 **Graph representation** - An SDRS can be seen as a directed acyclic graph where each DU is a vertex

0 **Directed labelled edges** for rhetorical relations

0 **Directed unlabelled edges** link CDUs to their content.

0 **Two kind of relations:**

0 **Coordinating**, e.g. Result, Narration (drawn horizontally)

0 **Subordinating**, e.g. Elaboration, Explanation, Contrast (drawn vertically)
Principles of natural selection. The theory of natural selection, such as it was initially described by Charles Darwin, is based on three principles: 1. Principle of variation 2. Principle of adaptation 3. Principle of heredity.

Scopes here can be described as: $C.(\pi_4, \pi_5) \land C.(\pi_5, \pi_6) \land Elab.(\pi_3, \pi_2) \land e \land Elab.(\pi_3, \pi_2) \land Elab(\pi_3, [\pi_4, \pi_5, \pi_6]) \land Elab(\pi_1, [\pi_3, \pi_4, \pi_5, \pi_6])$. 
0 **Right frontier constraint (RFC)** - A discourse constituent must be attached on the right frontier of the ongoing discourse.

Given a tree or a graph, an attachment cannot jump to a constituent on the left of the current one.

!!Claimed to be intuitive!!

Let $\Delta$ be a discourse structure with $\alpha$ the current (= last) constituent. A new constituent $\beta$ can be attached by a DR to $\gamma$ in $\Delta$ only if:

1. $\gamma = \alpha$, or
2. $\alpha$ is subordinated to $\gamma$.

a. John had a great evening last night.
 b. He had a great meal.
 c. He ate salmon.
 d. He devoured lots of cheese.
 e. He then won a dancing competition.
 f. It (the salmon) was a beautiful pink.
RFC requires that new discourse constituents cannot be integrated anywhere in the tree representing preceding discourse but only in a well-defined area situated at the Right Frontier of the tree.

- The constituents situated in this area are said to be *open* for attachment while all the others are said to be *closed* (Polanyi, 1988; Asher & Lascarides, 2003).
RST
tries to capture the intensions of the author as they are judged by the human annotator

SDRT
based on formal semantics

PDTB
based on lexically grounded relations

long-distance attachments are not possible

long-distance attachments are possible

long-distance attachments are possible
SDRT corpora

0 DISCOR
Brian Reese, Julie Hunter, Nicholas Asher, Pascal Denis and Jason Baldridge. Reference Manual for the Analysis and Annotation of Rhetorical Structure. 2007

0 ANNODIS

0 GEOPO
Lydia-Mai Ho-Dac : La position initiale dans l’organisation du discours : une exploration en corpus, thèse de doctorat, Université de Toulouse-Le Mirail, novembre 2007

0 STAC
DISCOR (Discourse Structure and Coreference Resolution)

0 Goal: to test hypotheses about the interaction between discourse structure and the resolution of anaphoric links.

0 Data: MUC and ACE corpora (because already have coreference)

0 14 relations
(cf. 78 in RST corpus)
ca. clauses, but not if embedded, e.g. do not segment wh-clause in
Privately held Arnold, which had about $750 million in billings and $90.7 million in revenue last year, handles advertising for such major corporations as McDonald’s Corp., ...

non-restrictive relative clauses introduce EDUs, unless doing so results in a discontinuous EDU

Many, but certainly not all, cases of syntactic subordination introduce a new elementary discourse unit. Complements of verbs of communication, for example, introduce EDUs (say, note, announce, etc).

appositions are EDUs

---

ANNODIS - (ANNOtation DIScursive)

- project 2007-2010 – texts annotation and parsing
- people: Nicholas Asher, Farah Benamara, Philippe Muller, Laure Vieu, Stergos Afantenos, etc.
- an annotated corpus of French written texts (news, wikipedia, linguistic papers, reports) for the study of discourse organization – 86 stories, 687,000 words,
- 3 years of annotation
- 3188 Elementary Discourse Units (EDU) and 1395 Complex Discourse Units (CDU) linked by 3355 rhetorical relations (e.g. contrast, elaboration, result, attribution, etc.)
- text annotated with various discourse phenomena
  - bottom-up approach: applying a compositional and logical model of discourse organization (SDRT)
  - top-down approach: starts from the text as a whole and focuses on the identification of configurations of cues signalling higher-level text segments, in an attempt to address the interplay of continuity and discontinuity within discourse (annotation of Enumerative Structures and Topical Chains)
- annotation tool Glozz (will be shown by the STAC corpus)
Rhetorical Relations

French

- 17 relations
- were chosen because they are more or less common to all the theories of discourse, or correspond to well-defined subgroups in fine-grained theories
- The intermediate level of granularity was chosen as a compromise between informativeness and reliability of the annotation process. It corresponds to the level chosen in the PDTB, and a coarse-grained RST.

SDRT corpora - STAC

0 running project: 2011-2016
0 ca 1100 negotiation dialogues, short texts (in English!) – much larger corpus than ANNODIS:
0 refined annotation tool Glozz
0 inter-annotator agreement better than for ANNODIS (basically because turns are annotated as EDU without further division)
domain level acts
offer: I'll give you 2 clay for a rock
counteroffer: How about 2 clay for a wheat?
accept: OK, it's a deal.
refusal: I don't think so.
has-resource: I have wheat
strategic comment: joel fancies a bit of your clay
other (non relevant for negotiation)

- relational rhetorical annotation

non-treelike structures certainly exist
long distance crossing dependencies
reflected in parsers output

Two-level annotation:

- segmentation of dialog turns into discourse units
- labelling with domain-related speech acts (negotiation moves)
- relational rhetorical annotation familiar from ANNODIS but with relations for dialogue.

• non-treelike structures certainly exist
• long distance crossing dependencies

STAC annotation

STAC

STAC 83

234 gotwood4sheep anyone got wheat for a sheep?
235 inca sorry, not me
236 CheshireCatGrin nope. you seem to have lots of sheep!
237 gotwood4sheep yup baaa
238 dmm i think i'd rather hang on to my wheat i'm afraid
239 gotwood4sheep kk I'll take my chances then...


Asher, Nicholas. Discourse Parsing for Text and Dialogue [2013]. MELODI. IRIT, CNRS

Asher, Nicholas (with Antoine Venant, Philippe Muller, Pascal Denis, Stergos Afantenos, Farah Benamara). Annotating and learning models of discourse structure [2015]

Asher, Nicholas and Lascarides, Alex (1997). Bridging. *Journal of Semantics*


Jacques Jayez. [2008]. The Right Frontier Constraint. ENS-LSH and L 2 C2


Lascarides, A. Semantics and Pragmatics of NLP Segmented Discourse Representation Theory. SPNLP. (slides)

Nicholas Asher; Laure Vieu. Subordinating and coordinating discourse relations. Lingua 115 (2005), 591–610.

Philippe Muller, S. Afantenos, Pascal D., N, Asher. Constrained decoding for text-level discourse parsing [2012], Proceedings of COLING 2012: Technical Papers, pages 1883–1900,


Brian Reese, Julie Hunter, Nicholas Asher, Pascal Denis and Jason Baldridge. [2007] DISCOR: Reference Manual for the Analysis and Annotation of Rhetorical Structure (Version 1.0)

SDRT done!
THEORIES & CORPORA

0 Halliday and Hasan – Cohesion in English
0 RST – Rhetorical Structure Theory
0 Discourse Graphbank
0 PDTB – Penn Discourse Treebank
0 SDRT - Segmented Discourse Representation Theory
0 CCR – Cognitive approach to coherence relations
CCR (Cognitive approach to coherence relations)

Coherence relations are considered as cognitive entities (coherence relations and their linguistic marking affect the cognitive representation of discourse, e.g. text understanding, they are psychological entities rather than merely an analytic tool).

Proved by series of psycholinguistic experiments (Spooren, 1989; Haberlandt, 1982; Sanders 1986, e.g. linguistic marking appears to lead to faster processing of the following discourse segment).

CCR (Sanders et al.)
Investigating coherence relations

- Relational account
  - Focus on the meaning of the relation and not on the meaning of each specific segment
  - Classification in terms of cognitive primitives

- Language-based account
  - Focus not so much on relations themselves but rather on a study of linguistic devices that are used to signal relations explicitly
Although he worked hard, he failed the exam.

CCR (Sanders et al.)

Four cognitive primitives

**Although he worked hard, he failed the exam.**

- discourse segment $S_1$
  - expresses $P$
- discourse segment $S_2$
  - expresses $Q$
CCR (Sanders et al.)

Four cognitive primitives

0 basic operation (Additive – Temporal – Causal / Conditional)

originally (Sanders et al. 1992) only additive (weakly connected) vs. causal (strongly connected) were distinguished

She got wet and her friend got wet too.

There was a lot of rain. Later, storms came in.

She got wet because it rained.
CCR (Sanders et al.)

Four cognitive primitives

0 basic operation (Additive – Temporal – Causal / Conditional)

0 source of coherence (semantic/ pragmatic)

She got wet because it rained.

A relation is objective (semantic) when both segments happen in the real world and are thus facts. The speaker’s opinion is not reflected in the relation.

She is not at home, because her car is not there

A relation is subjective (pragmatic) if one or both segments express an opinion, argument, claim or conclusion of the speaker.
CCR (Sanders et al.)
Four cognitive primitives

0 basic operation (Additive – Temporal – Causal / Conditional)
0 source of coherence (semantic/pragmatic)
0 order of segments (basic/non-basic)

Although he worked hard, he failed the exam.

A relation with a basic order has an antecedent, followed by a consequent. The antecedent is the cause of the argument, the consequent is the consequence or the claim.

She got wet because it rained.

A relation with a non-basic order has a consequent which precedes the antecedent. The cause or the claim thus precedes the cause of the argument.
CCR (Sanders et al.)
Four cognitive primitives

0 basic operation (Additive – Temporal – Causal / Conditional)
0 source of coherence (semantic/pragmatic)
0 order of segments (basic/non-basic)
0 polarity (positive/negative)

She got wet because it rained.

\[ S_1 = Q \quad S_2 = P \]

A relation is positive if the two discourse segments \( S_1 \) and \( S_2 \) function in the basic operation as antecedent (P) and consequent (Q) respectively.

Although he worked hard, he failed the exam.

not-\( S_1 = P \) \quad not-\( S_2 = Q \)

A relation is negative if not \( S_1 \) and \( S_2 \), but their negative counterparts, not-\( S_1 \) and not-\( S_2 \), function in the basic operation.
DiscAn corpora

Towards a discourse annotation system for Dutch language corpora

0 DiscAn project – integrating existing corpora of Dutch discourse phenomena in the CLARIN infrastructure
  0 set of corpus analyses has been standardized (both in terms of raw data -- the texts -- and analyses) and opened up for further scientific research
0 text have been annotated for discourse phenomena during last ca 15 years at several universities in the Netherlands and Belgium
0 format: Excel tables, doc files, SPSS files etc. → brat?
0 visualization ANNIS

DiscAn subcorpora

0 Degand 2.2: compiled for a study of the causal connectives "aangezien", "want" and "omdat" in Dutch news, 143 cases from a Dutch newspaper (NRC Handelsblad from 1994)


0 PanderMaatSanders: causal connectives "daardoor", "darom" and "dus" in Dutch news

0 Persoon Corpus: causal connectives "want" and "omdat" in Dutch spontaneous conversations, from the Corpus of Spoken Dutch


0 SandersSpoooren Corpus: causal connectives "want" and "omdat" in several types of discourse
## DiscAn subcorpora

<table>
<thead>
<tr>
<th>Discourse phenomena</th>
<th>Author</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causal connectives</td>
<td>Bekker (2006)</td>
<td>500 explicit (<code>doordat</code>, <code>want</code>, <code>dus</code>, <code>daarom</code>, <code>nadat</code>, <code>voordat</code>) / 200 implicit</td>
</tr>
<tr>
<td>Causal connectives</td>
<td>Degand (2001)</td>
<td>150 (<code>want</code>, <code>aangezien</code>, <code>ondat</code>) from newspapers</td>
</tr>
<tr>
<td>Coherence relations</td>
<td>Den Ouden (2004)</td>
<td>70 (causal implicit, non-causal)</td>
</tr>
<tr>
<td>Connectives</td>
<td>Evers-Vermeul (2005)</td>
<td>600 historical data / 4400 from Childes</td>
</tr>
<tr>
<td>Causal connectives</td>
<td>Pander Maat &amp; Degand (2001)</td>
<td>150 (<code>dus</code>, <code>daarom</code>) from newspaper corpora</td>
</tr>
<tr>
<td>Causal connectives</td>
<td>Pander Maat &amp; Sanders (2000)</td>
<td>150 (<code>dus</code>, <code>daarom</code>, <code>daardoor</code>) from a newspaper-corpus (Volkskrant)</td>
</tr>
<tr>
<td>Causal connectives</td>
<td>Persoon (2010)</td>
<td>105 (<code>ondat</code>, <code>want</code>) from CGN</td>
</tr>
<tr>
<td>Causal connectives</td>
<td>Pit (2003)</td>
<td>200 (<code>aangezien</code>, <code>ondat</code>, <code>doordat</code>, <code>want</code>) newspaper / 100 (<code>ondat</code>, <code>doordat</code>, <code>want</code>) narrative; from newspaper (Volkskrant) and fictional books</td>
</tr>
<tr>
<td>Causal connectives</td>
<td>Sanders &amp; Spooren (2009)</td>
<td>100 newspaper (Volkskrant) / 275 from CGN / 80 from Chat (<code>want</code>, <code>ondat</code>)</td>
</tr>
<tr>
<td>Coherence relations</td>
<td>Sanders &amp; van Wijk (1996)</td>
<td>100 children's explanatory texts; ca. 1500 coherence relations</td>
</tr>
<tr>
<td>Coherence relations</td>
<td>Spooren &amp; Sanders (2008)</td>
<td>1100 coherence relations (children elicit responses)</td>
</tr>
<tr>
<td>Causal connectives</td>
<td>Spooren et al. (2010)</td>
<td>275 (<code>want</code>, <code>ondat</code>) spoken, from CGN; 100 (<code>want</code>, <code>ondat</code>) written</td>
</tr>
<tr>
<td>Causal connectives</td>
<td>Stukker (2005)</td>
<td>300 (<code>daardoor</code>, <code>daarom</code>, <code>dus</code>) newspaper / 300 historical data (<code>daarom</code>, <code>dus</code>)</td>
</tr>
<tr>
<td>Coherence relations</td>
<td>Vis (2011)</td>
<td>135 texts; 643 subjective relations</td>
</tr>
<tr>
<td>Connectives</td>
<td>Van Veen (2011)</td>
<td>1951 <code>waarom</code> (why-) questions &amp; their answers (Childes)</td>
</tr>
</tbody>
</table>

From Sanders et al. 2012
DiscAn Annotation

Four annotated categories:

0 Polarity (negative/positive)

0 Basic operation

0 Source of coherence (objective/subjective)

0 Order

BUT! Temporal relations can have a chronological order (basic order) and non-chronological (non-basic). Two events can also happen simultaneously. In this case, order is not annotated.

BUT! Order for additive relations is not marked, because they are symmetric.

BUT! Temporal and non-causal negative relations do not differ in source of coherence, because they are an objective representation of reality by nature.


CCR done!
THEORIES & CORPORA

0 Halliday and Hasan – Cohesion in English
0 RST – Rhetorical Structure Theory
0 SDRT - Segmented Discourse Representation Theory
0 PDTB – Penn Discourse Treebank and PDTB-like
0 Discourse Graphbank
0 CCR – Cognitive approach to coherence relations
0 How to compare the theories/approaches/structures/corpora...?
0 Can we translate from one corpus to another, thus extending the range of data available for performing automated tasks?
0 Once you choose one of these theories you get the corpus that looks completely different than another corpus in another theory. (RST corpus looks completely different than PDTB)
Comparative attempts


- Anja Nedoluzhko, Ekaterina Lapshinova, Kerstin Kunz: PDT vs. GECCo (2015: LAW-IX NAACL, TextLink meetings in Louvain-la-Neuve and Saarbrucken, LREC subm.)


Creating common standard

- ISO standard for the annotation of semantic relations in discourse

- applying Sanders’ dimensions (April 2015, Fribourg)
Comparative attempts


- Anja Nedoluzhko, Ekaterina Lapshinova, Kerstin Kunz: PDT vs. GECCo (2015: LAW-IX NAACL, TextLink meetings in Louvain-la-Neuve and Saarbrucken, LREC subm.)


Creating common standard

- ISO standard for the annotation of semantic relations in discourse
- applying Sanders’ dimensions (April 2015, Fribourg)
**Anja Nedoluzhko, Ekaterina Lapshinova, Kerstin Kunz:** **PDT vs. GECCo** (2015)

**PDT vs. GECCo**

- DSDs
- ellipsis
- bridging/lexical cohesion
- coreference

### Discourse Markers (attitude markers, modal particles)
- in PDT not considered as connectives

**Prague Dependency Treebank (Prague)**

**German-English Contrasts in Cohesion (Saarland Uni)**
Comparative attempts


- Anja Nedoluzhko, Ekaterina Lapshinova, Kerstin Kunz: PDT vs. GECCo (2015: LAW-IX NAACL, TextLink meetings in Louvain-la-Neuve and Saarbrucken, LREC subm.)


Creating common standard

- ISO standard for the annotation of semantic relations in discourse applying Sanders’ dimensions (April 2015, Fribourg)
0 Use CCR as an intermediate language

<table>
<thead>
<tr>
<th>Basic Operation</th>
<th>Source of Coherence</th>
<th>Order</th>
<th>Polarity</th>
<th>Class</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causal</td>
<td>Semantic</td>
<td>Basic</td>
<td>Positive</td>
<td>1.</td>
<td>Cause–consequence</td>
</tr>
<tr>
<td>Causal</td>
<td>Semantic</td>
<td>Basic</td>
<td>Negative</td>
<td>2.</td>
<td>Contrastive cause–consequence</td>
</tr>
<tr>
<td>Causal</td>
<td>Semantic</td>
<td>Nonbasic</td>
<td>Negative</td>
<td>4.</td>
<td>Contrastive consequence–cause</td>
</tr>
<tr>
<td>Causal</td>
<td>Pragmatic</td>
<td>Basic</td>
<td>Positive</td>
<td>5a.</td>
<td>Argument–claim</td>
</tr>
<tr>
<td>Causal</td>
<td>Pragmatic</td>
<td>Basic</td>
<td>Negative</td>
<td>5b.</td>
<td>Instrument–goal</td>
</tr>
<tr>
<td>Causal</td>
<td>Pragmatic</td>
<td>Nonbasic</td>
<td>Positive</td>
<td>5c.</td>
<td>Condition–consequence</td>
</tr>
<tr>
<td>Causal</td>
<td>Pragmatic</td>
<td>Nonbasic</td>
<td>Negative</td>
<td>5d.</td>
<td>Contrastive argument–claim</td>
</tr>
<tr>
<td>Causal</td>
<td>Pragmatic</td>
<td>Nonbasic</td>
<td>Positive</td>
<td>6.</td>
<td>Claim–argument</td>
</tr>
<tr>
<td>Causal</td>
<td>Pragmatic</td>
<td>Nonbasic</td>
<td>Negative</td>
<td>7a.</td>
<td>Goal–instrument</td>
</tr>
<tr>
<td>Causal</td>
<td>Pragmatic</td>
<td>Nonbasic</td>
<td>Negative</td>
<td>7b.</td>
<td>Consequence–condition</td>
</tr>
<tr>
<td>Causal</td>
<td>Pragmatic</td>
<td>Nonbasic</td>
<td>Negative</td>
<td>7c.</td>
<td>Contrastive claim–argument</td>
</tr>
<tr>
<td>Additive</td>
<td>Semantic</td>
<td>—</td>
<td>Positive</td>
<td>9.</td>
<td>List</td>
</tr>
<tr>
<td>Additive</td>
<td>Semantic</td>
<td>—</td>
<td>Negative</td>
<td>10a.</td>
<td>Exception</td>
</tr>
<tr>
<td>Additive</td>
<td>Pragmatic</td>
<td>—</td>
<td>Positive</td>
<td>10b.</td>
<td>Opposition</td>
</tr>
<tr>
<td>Additive</td>
<td>Pragmatic</td>
<td>—</td>
<td>Negative</td>
<td>11.</td>
<td>Enumeration</td>
</tr>
<tr>
<td>Additive</td>
<td>Pragmatic</td>
<td>—</td>
<td>Negative</td>
<td>12.</td>
<td>Concession</td>
</tr>
</tbody>
</table>

CCR – RST mapping

<table>
<thead>
<tr>
<th>Basic op.</th>
<th>Source of co.</th>
<th>Order</th>
<th>Polarity</th>
<th>CCR Relation</th>
<th>Additional criteria</th>
<th>RST Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causal</td>
<td>Objective</td>
<td>Basic</td>
<td>Positive</td>
<td>Cause–consequence</td>
<td>conditional</td>
<td>Violation cause/result</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

from Sanders et al. 1992

from Sanders et al. 2015
Ted Sanders, Let's try to make annotation systems communicate – **towards a systematic approach of coherence relations**. Fribourg, 2015.

<table>
<thead>
<tr>
<th>Example 1:</th>
<th>Example 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>The door slammed because there is strong wind outside.</td>
<td>Max is a very good skier, because he won the competition twice last year.</td>
</tr>
<tr>
<td>➢ RST: Non-Volitional cause / Explanation ?</td>
<td>➢ RST: Evidence ?</td>
</tr>
<tr>
<td>➢ PDTB: (CONTINGENCY.Cause.) reason ?</td>
<td>➢ PDTB: (CONTINGENCY.) Pragmatic cause / justification ?</td>
</tr>
<tr>
<td>➢ CCR: positive, causal, objective, non-basic (Consequence-cause)</td>
<td>➢ CCR: Positive, causal, subjective, non-basic (Claim-argument)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 3:</th>
<th>Example 4:</th>
</tr>
</thead>
<tbody>
<tr>
<td>John is tall but Fred is small.</td>
<td>Jane married Mark even though she does not love him.</td>
</tr>
<tr>
<td>➢ RST: Contrast ?</td>
<td>➢ RST: Antithesis ?</td>
</tr>
<tr>
<td>➢ PDTB: (COMPARISON.Contrast.)opposition ?</td>
<td>➢ PDTB: (COMPARISON.Concession.) contra-expectation ?</td>
</tr>
<tr>
<td>➢ CCR: Negative, additive, objective (opposition)</td>
<td>➢ CCR: negative, causal, objective, non-basic (Contrastive consequence-cause)</td>
</tr>
</tbody>
</table>

*from Sanders, 2015*

0 **idea**: Use CCR (Cognitive approach to Coherence Relations, Sanders, Spooren & Noordman 1992) as an intermediate language between different frameworks.

0 **question**: To what extent can PDTB relations be analysed consistently using CCR dimensions?

0 2 annotators analyzed 1197 relations independently using PDTB 3.0 and CCR, respectively.

0 Annotations mapped onto each other to investigate consistency of relation meanings across theories.
Overall, 69% of the PDTB relations were consistently categorized as belonging to the target CCR class.

Analysis of random sample of 50 disagreements: 48% of disagreements due to differences between the theories.

Other disagreements due to difference in segmentation or interpretation of relation (14%) and to annotation errors (38%)

### Examples of differences between theories

- The connective ‘but’ indicates a negative relation in CCR, but not necessarily in PDTB.
  
  (1) *She’s by a Northern-based sire.* [implicit but] *I think he’s dead now perhaps.*

- Argumentative relations classified as causal in CCR, but additive ‘Expansion’ in PDTB.
  
  (2) *I used the weight room facility for exercising.* [impl. because] *I exercise from physiotherapy that I had to do.*
Comparative attempts


0 Anja Nedoluzhko, Ekaterina Lapshinova, Kerstin Kunz: PDT vs. GECCo (2015: LAW-IX NAACL, TextLink meetings in Louvain-la-Neuve and Saarbrucken, LREC subm.)

0 Venant –Asher – Muller – Pascal – Afantenos: Expressivity and comparison of models of discourse structures, Sigdial 2013.

Creating common standard

0 ISO standard for the annotation of semantic relations in discourse applying Sanders’ dimensions (April 2015, Fribourg)
Interprovincial Pipe Line Co. said it will delay a proposed two-step, 830 million dollar [(US$705.6 million)] expansion of its system because Canada’s output of crude oil is shrinking.

Mixed Nuclearity Principle
- NS relations only transmit nucleus argument to a parent relation.
- Restatement(π2, π3) ∧ Explanation(π2, π4) ∧ Attribution(π1, π2)

If RST is interpreted indirectly, we should be able to describe how the unpacking of the real arguments in the tree structure works!!!
Comparisons require a language expressive enough to express semantic scopes for all discourse theories and interpretative mappings from the different structures into this language.

Venant et al. 2013 propose a language (formal semantic description) in which all theories can be described. Then we can look and compare the commonalities and restrictions, incl. decoding one from the other.

Immediate interpretation of a RS-Tree can be decoded back to a SDRS (it gives lots of CDUs).

Relaxed Nuclearity Principle $\rightarrow$ RS Tree can be decoded into a set of SDRSs.

The same with dependency trees.

DP and RST-trees are both underspecified version of a fully specified structure 😊
Comparative attempts


0 Anja Nedoluzhko, Ekaterina Lapshinova, Kerstin Kunz: PDT vs. GECCo (2015: LAW-IX, NAACL, TextLink meetings in Louvain-la-Neuve and Saarbrucken, LREC subm.)

0 Venant –Asher – Muller – Pascal – Afantenos: Expressivity and comparison of models discourse structures, Sigdial 2013.

Creating common standard

0 ISO standard for the annotation of semantic relations in discourse applying Sanders’ dimensions (April 2015, Fribourg)
compared approaches

created an ISO, very close to current PDTB 3.0 version

- Types of objects connected by discourse relations,
- explicit connective/implicit/AltLex
- representation of discourse structure
- semantic description of discourse relations
- pragmatic variants of discourse relations
- hierarchical classification of discourse relations
- representation of (a)symmetry of relations
- relative importance of arguments for text meaning/structure
Constraints are placed on the semantic nature of arguments rather than on their syntactic form. An argument of a discourse relation must denote a certain type of abstract object. The ISO scheme remains neutral on this issue and does not specify any constraints on the extent or adjacency of argument realizations.
compared approaches created an ISO, very close to current PDTB 3.0 version

- Types of objects connected by discourse relations
  - explicit connective/implicit/AltLex

- representation of discourse structure
- semantic description of discourse relations
- pragmatic variants of discourse relations
- hierarchical classification of discourse relations
- representation of (a)symmetry of relations
- relative importance of arguments for text meaning/structure
Explicit connectives/Implicit/AltLex

- RST-signaling: all elements having a connecting function
- PDTB: explicit, implicit, AltLexes
- PDT: explicit
- DGB
- SDRT
- CCR: explicit, implicit
- H&H: explicit

as it is

ISO 24617-8
Compared approaches created an ISO, very close to current PDTB 3.0 version.

Types of objects connected by discourse relations, explicit connective/implicit/AltLex, representation of discourse structure, semantic description of discourse relations, pragmatic variants of discourse relations, hierarchical classification of discourse relations, representation of (a)symmetry of relations, relative importance of arguments for text meaning/structure.

ISO 24617-8
ISO takes a pre-theoretical stance involving low-level annotation of discourse; individual relations can then be annotated further to project a higher-level tree or graph structure, depending on one’s theoretical preferences.

ISO 24617-8

Relations along with their arguments are annotated without being combined to form a structure that encompasses the entire text.

PDTB

Relations along with their arguments are annotated without being combined to form a structure that encompasses the entire text.

DGB

Allow general graphs that allow multiple parents and crossing.

RST

A tree representation to subsume the complete text of the discourse.

SDRT

Allow directed acyclic graphs that allow for multiple parents, but not for crossing.

CCR

Relations along with their arguments are annotated without being combined to form a structure that encompasses the entire text.

H&H

Relations along with their arguments are annotated without being combined to form a structure that encompasses the entire text.
compared approaches
created an ISO, very close to
current PDTB 3.0 version

0 Types of objects connected by discourse relations, ✓
0 explicit connective/implicit/AltLex ✓
0 representation of discourse structure ✓
0 semantic description of discourse relations
0 pragmatic variants of discourse relations
0 hierarchical classification of discourse relations
0 representation of (a)symmetry of relations
0 relative importance of arguments for text meaning/structure
in 'informational' terms, i.e., in terms of the content of the arguments

RST
in terms of the intended effects on the hearer/reader

PDTB
in 'informational' terms, i.e., in terms of the content of the arguments

DGB
in 'informational' terms, i.e., in terms of the content of the arguments

SDRT
in 'informational' terms, i.e., in terms of the content of the arguments

CCR
in 'informational' terms, i.e., in terms of the content of the arguments

H&H
in 'informational' terms, i.e., in terms of the content of the arguments

ISO 24617-8

compared approaches
created an ISO, very close to current PDTB 3.0 version

- Types of objects connected by discourse relations
- Explicit connective/implicit/AltLex
- Representation of discourse structure
- Semantic description of discourse relations
- Pragmatic variants of discourse relations
- Hierarchical classification of discourse relations
- Representation of (a)symmetry of relations
- Relative importance of arguments for text meaning/structure

RST, DGB, PDTB, SDRT, CCR, H&H

ISO 24617-8
**Semantic-Pragmatic Distinction, for All Relation Types**

However, the ISO scheme does not encode this distinction on the relation, but on the arguments of the relation.

- **RST**
  - Semantic-pragmatic, for some relation types (volutional - non volutional) subject matter relations vs. presentational relations

- **DGB**
  - Do NOT distinguish pragmatic

- **SDRT**
  - Content -- metatalk

- **CCR**
  - Semantic-pragmatic distinction, for all relation types
  - Subjective – Objective

- **PDTB**
  - Semantic-pragmatic, for some relation types

- **ISO 24617-8**
  - Semantic-pragmatic distinction, for all relation types
  - However, the ISO scheme does not encode this distinction on the relation, but on the arguments of the relation

- **H&H**
  - Internal-external distinction

**Pragmatic Variants of Discourse Relations**

compared approaches created an ISO, very close to current PDTB 3.0 version

Types of objects connected by discourse relations, ✓
explicit connective/implicit/AltLex ✓
representation of discourse structure ✓
semantic description of discourse relations ✓
pragmatic variants of discourse relations ✓

hierarchical classification of discourse relations

representation of (a)symmetry of relations
relative importance of arguments for text meaning/structure
Hierarchical Classification of Discourse Relations

- **RST**: yes, 2 levels, big groups: Causal-conditional, Contrastive, Additive
- **PDTB**: yes, 3 levels
- **PDT**: yes, 2 levels
- **ISO 24617-8**: yes, 2 levels
- **DGB**: no levels?
- **SDRT**: no levels?
- **CCR**: no, 4 criteria + additional features
- **H&H**: no, only basic groups

A ‘flat’ set of core relations
compared approaches

created an ISO, very close to current PDTB 3.0 version

0 Types of objects connected by discourse relations
0 explicit connective/implicit/AltLex
0 representation of discourse structure
0 semantic description of discourse relations
0 pragmatic variants of discourse relations
0 hierarchical classification of discourse relations
0 representation of (a)symmetry of relations
0 relative importance of arguments for text meaning/structure
Argument spans are named Arg1 and Arg2 according to syntactic criteria, including linear order, and the asymmetrical relations are defined in terms of the Arg1 and Arg2 labels (for example, the relation Cause:Reason has Arg2 as the cause and Arg1 as the effect, while the relation Cause:Result has Arg1 as the cause and Arg2 as the effect.

The argument span ordering is one of the ‘cognitive’ primitives underlying the scheme. The relation Cause-Consequence captures the ‘basic’ order for the semantic causal relation, with the cause appearing before the effect, whereas the relation Consequence-Cause is used for the reversed order of the arguments.

Annotations abstract over the linear ordering for argument realizations, since this is not a semantic distinction. Instead, asymmetry is represented by specifying the argument roles in the definition of each relation.
compared approaches created an ISO, very close to current PDTB 3.0 version

- Types of objects connected by discourse relations
- explicit connective/implicit/AltLex
- representation of discourse structure
- semantic description of discourse relations
- pragmatic variants of discourse relations
- hierarchical classification of discourse relations
- representation of (a)symmetry of relations
- relative importance of arguments for text meaning/structure

RST, DGB, PDTB, SDRT, CCR, H&H

ISO 24617-8
the relative role of arguments for the text (meaning or structure) as a whole is not represented directly

Argument spans are named Arg1 and Arg2 according to syntactic criteria, including linear order, and the asymmetrical relations are defined in terms of the Arg1 and Arg2 labels (for example, the relation Cause:Reason has Arg2 as the cause and Arg1 as the effect, while the relation Cause:Result has Arg1 as the cause and Arg2 as the effect.)

yes, classifies a relation as ‘subordinating’ or ‘coordinating’, depending on what structural configuration the arguments create in the discourse graph (Asher and Vieu, 2005)

yes, ‘dominance’, deriving a single assertion from a discourse relation that connects two segments, and distinguishing relations in terms of how this single assertion should be derived. In subordinating relations, in particular, the assertion associated with the relation is obtained from the ‘dominant’ segment, as specified in the relation definitions.
How many discourse functions can a single unit have?

*We bought the apartment, but then we rented it out. RST – one, PDTB-like, SDRT, DGB: both, other theories?*

Can substructures of discourse serve as arguments of coherence relations? (e.g. yes for SDRT)

Attachment principles - where can we attach a new information:

- attachment to the last in a queue, to a preceding sentence (dynamic semantics)
- long distance dependencies (RST-treebank, SDRT - prove that there are long-distance dependencies: 35-40% of the attachments are outside the sentence and beyond the previous sentence) - there are some principles how you can attach, graph, positions of the graph, primary attachment portions (right frontier constraint)
- dialogues give interesting constraints (to backwards attachment - Although I’m tired, I’m enjoying it here)
Is that all? Of course not, but ...

Thank you for attention!
Acknowledgements

I would like to thank my colleagues, whose materials published in articles and slides I used: Nicholas Asher, Vera Demberg, Markus Egg, Annemarie Friedrich, Jacques Jayez, Alex Lascarides, Lynn Carlson, Daniel Marcu, Gisela Redeker, Ines Rehbein, Ted Sanders, Merel Scholman, Manfrede Stede, Maite Taboada, etc.

Many thanks to my discourse colleagues (Eva Hajičová, Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Jiří Mírovský, etc.)

The research was supported from the Grant Agency of Czech Republic (grant GA 16-05394S, Structure of coreferential chains in parallel language data)