How complex is discourse structure?

Markus Egg and Gisela Redeker
Humboldt-Universität Berlin/Rijksuniversiteit Groningen

LREC 2010
University of Malta, 20 May, 2010
Outline of the talk

• introduction: representations of discourse structure

• crucial phenomena
  – crossed dependencies
  – multiple-parent structures
  – a combination of these: potential list structures

• conclusion and outlook
Introduction 1

- discourse is structured by discourse relations that combine smaller segments into larger ones
- discourse relations typically comprise cause/result, lists, or elaboration
- most discourse structure theories and annotated corpora assume that discourse structure is a tree
- in particular those that implement some version of Rhetorical Structure Theory (RST; Mann and Thompson 1988; Taboada and Mann 2006)
  - the WSJ Discourse Tree Bank (Carlson et al. 2003)
  - the Potsdam Commentary Corpus (Stede 2004)
- this assumption has come under attack as too restricted (Wolf and Gibson 2005, 2006; Lee et al. 2008)
• Wolf and Gibson (W&G) claim that discourse structure is much more complex and requires a representation in terms of chain graphs

(1) \((C_1)\) “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)
but the discourse structure of (1) can also be modelled as tree (Egg and Redeker 2008)

(3)\[ \text{elab}_n \quad \text{attr}_n \quad C_4 \]

\[ \text{elab}_n \quad C_3 \]

\[ C_1 \quad C_2 \]
Introduction 4

- such competing analyses of the examples suggest evaluating W&G’s corpus
  - the *Discourse Graphbank* (DGB; Wolf et al. 2005)
  - 135 texts from the AP Newswire and Wall Street Journal
- it comprises 10.3% more relations than a tree analysis could maximally have
- there are crossed dependencies
- 41.22% of the segments have multiple parents (W&G 2005)
- our goal: distinguish the complexity inherent in the data and the one arising from specific design choices in W&G’s annotation
- our sample: the first 14 texts in the DGB (approx. 10% of the corpus)
Crossed dependencies

- crossed dependencies in the DGB
  - relations link (widely) non-adjacent discourse segments
  - many of these relations are **ELABORATION** relations
    * 50.5% of crossed dependencies in the DGB are **ELABORATION**
    * in our sample, this holds for 69% of the relations with a gap of $\geq 6$ units

- **ELABORATION** relations are problematic anyway (e.g., Knott et al. 2001)
  - many of them operate **between coherence and cohesion**
  - they target concepts and not entire discourse segments
  - they appear to be inspired by lexical or referential cohesion

- correlation between two problems in the DGB
  - relations that are based on cohesion (Egg and Redeker 2008)
  - relations that introduce crossed dependencies (Webber et al. 2003)
Multiple-parent structures 1

• a typical instance of multiple-parent structures (MPS) in the DGB: embedded quotes, as in (4) [= (1)]

(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)

• these texts very often quote a source
  – message and source are linked by \textsc{Attribution} (Carlson and Marcu 2001)
  – the message is considered \textbf{more important} than the source
  – importance is modelled in terms of \textsc{subordination}
  – the source is encoded as \textsc{satellite} and the message as \textsc{nucleus}
Multiple-parent structures 2

- the critical instances have the source embedded in the message
- for embedded sources, W&G annotate the attribution to left and right and link parts of the message pairwise
- example (4) in their analysis [= (2)]
Multiple-parent structures 3

- RST-based analysis of (4)

\[
(5) \; [\; (= \; (3)] \\
\text{elab}_n \\
\text{attr}_n \; \text{C}_4 \\
\text{elab}_n \; \text{C}_3 \\
\text{C}_1 \; \text{C}_2
\]

- this analysis uses the **nuclearity principle** of Marcu (1996)
- the RST-based analyses have one **ATTRIBUTION** relation less
- the sample comprises 11 such embedded-source constellations
- these additional relations are 8% of the 138 excess relations for the sample
- this is approx. 1/3 of MPS in general, further work is necessary
**Multiple-parent structures 4**

- Lee et al. (2008) annotate MPS in the Penn Discourse Treebank (PDTB)
  
  (6) *If this seems like pretty weak stuff around which to raise the protectionist barriers,* \((C_1)\) it may be \((C_2)\) because these shows need all the protection they can get. \((C_3)\) European programs usually target only their own local audience \(\ldots\). (2361)

- in (6), they regard \(C_2\) as the immediate argument of two causal discourse relations, linking it to both \(C_1\) and \(C_3\)

- empirical evidence:
  - each discourse relation and its arguments are annotated independently
  - in cases like (6), a (syntactically) subordinated segment is reselected
  - there are 349 instances of this constellation in the PDTB
Multiple-parent structures 5

- in an alternative tree-structure analysis of (6), the causal relation introduced by *because* links $C_1$ to the segment consisting of $C_2$ and $C_3$

- general question: relation between Lee et al.’s (2009) results and the PDTB annotation manual (Prasad et al. 2006)
  - annotators were explicitly required to specify the smallest arguments possible for the discourse relation in question
  - many satellites can be left out in a text without resulting in discoherence
  - in (6), this might have caused the annotators to choose $C_2$ (instead of $C_2$ and $C_3$) as the second argument of *because*
  - manual investigation of at least a relevant sample of the examples needed
Potential list structures 1

- multiple attachments and crossed dependencies also show up in potential list structures
  - they are of the form ‘$A B_1 B_2 \ldots B_n$’
  - all $B_i$ stand in the same relation $Rel$ to $A$
  - all $B_i$ could be interpreted as list (or sequence)

- in (7), $C_1$ is elaborated by $[C_2 C_3]$, $C_4$, and $C_5$

(7) $(C_1)$ Students learn to program a computer and automated machines linked to it in a complete manufacturing operation $(C_2)$ retrieving raw materials from the storage shelf unit $(C_3)$ which can be programmed to supply appropriate parts from its inventory; $(C_4)$ lifting and placing the parts in position with the robot’s arm; $(C_5)$ and shaping parts into finished products at the lathe. (ap-890101-0002)
Potential list structures 2

• W&G analyse these cases in that
  – each $B_i$ is linked to $A$ by $Rel$ individually
  – the $B_i$ are linked by parallelism (or elaboration)

• example (7) in their analysis
Potential list structures 3

• an RST-based analysis of (7) first combines the $B_i$ and links them to $A$ in one go

(8)

```
      elab_n
     /   \
list /     \
  /       \
C_1  elab_n  C_4  C_5  \\
 /       /   \
C_2  C_3  C_4  C_5
```

• W&G obtain many additional relations in this way

• their annotation manual requires annotators to integrate new material in a non-hierarchical way

• in our corpus sample there are five of these cases with three list elements each

• this accounts for 15 (10.9%) of the problematic relations

Markus Egg and Gisela Redeker, LREC 2010
Conclusion and outlook

• we evaluated claims that discourse structure is more complex than tree structures

• there seems to be an interdependence between annotation manuals and the resulting complexity of representations of discourse structure

• we identified a number of crucial potentially non-treelike discourse constellations for which alternative tree-structure analyses are feasible

• it is the subject of further research to investigate whether this holds for all potentially non-treelike structures
References


Markus Egg and Gisela Redeker, LREC 2010
Multiple-parent structures 3

- RST-based analysis of (4)

(9) [= (5)]

```
  elab_n
   / \  
  attr_n C_4
   / \  
 elab_n C_3
   / \  
 C_1  C_2
```

- this analysis uses the **nuclearity principle** (Marcu 1996):
  A relation between a complex segment $A$ and another segment $B$ implies the same relation between the nucleus of $A$, and $B$
  - in (3), the ELABORATION between $C_1$-$C_3$ and $C_4$ is based on the same relation between $C_1$-$C_2$ (the nucleus of $C_1$-$C_3$) and $C_4$
  - the source $C_3$ is not a right boundary for the information
  - $C_3$ can indicate the source for $C_4$, too