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

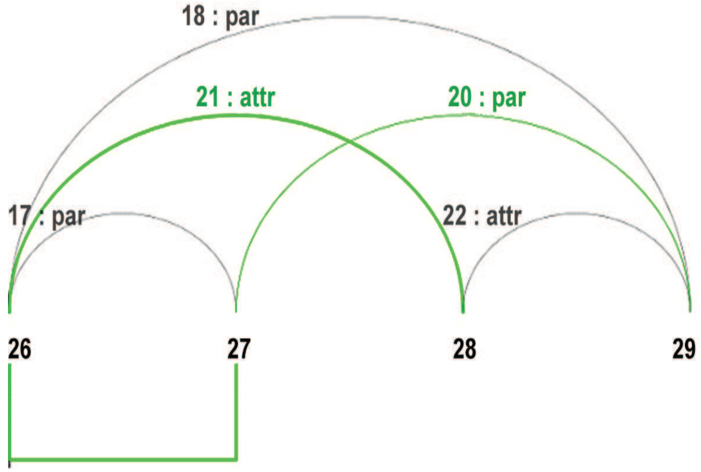
Introduction 1

- discourse is structuctured 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)

Introduction 2

- 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)

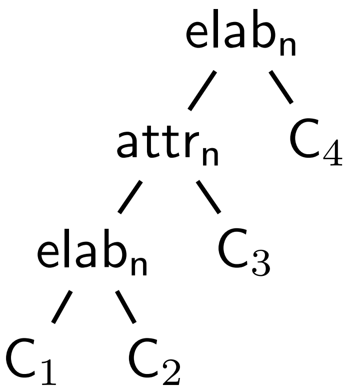


(2)

Introduction 3

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

(3)



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 beween 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

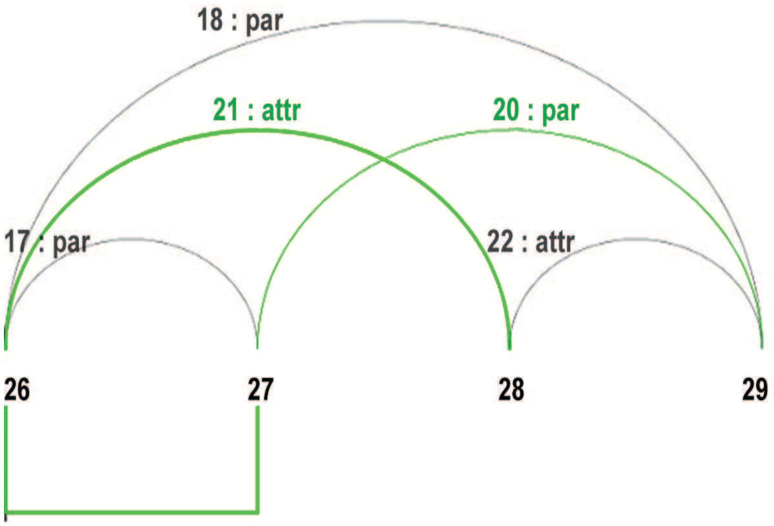
- a typical instance of multiple-parent structures (MPS) in the DGB: embedded quotes, as in (4) [= (1)]

  (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 ATTRIBUTION (Carlson and Marcu 2001)
  - the message is considered more important than the source
  - importance is modelled in terms of subordination
  - the source is encoded as satellite and the message as nucleus

Multiple-parent structures 2

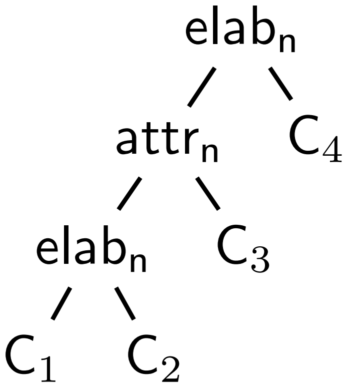
- the critical instances have the source embedded in the message
- for embedded sources, W&G annotate the attribution to left and right and link parts of the message pairwise
- example (4) in their analysis [= (2)]



Multiple-parent structures 3

- RST-based analysis of (4)

(5) [= (3)]



- 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 (. . .).* (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

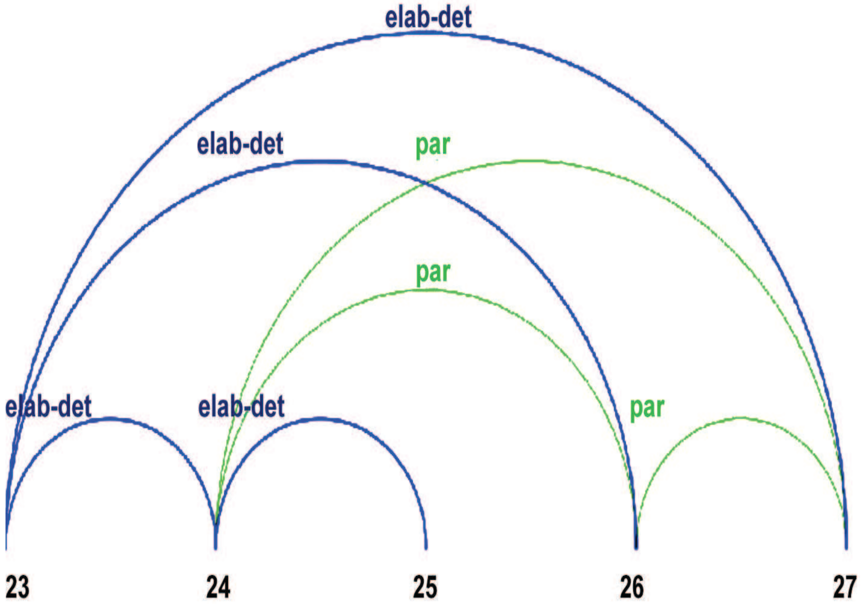
Potential list structures 1

- 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

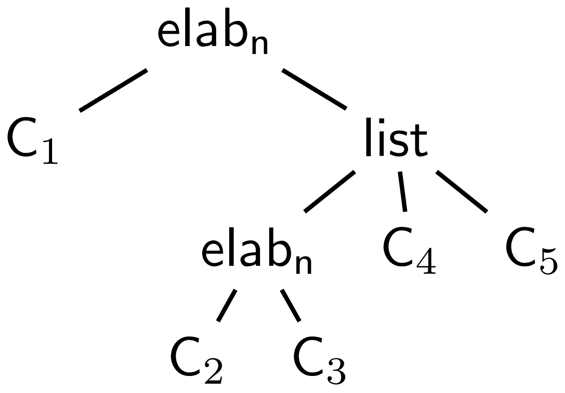
- W&G analyse these cases in that
  - each B_i is linked to A by Rel individually
  - the B_i are linked by parallelism (or elaboration)
- example (7) in their analysis



Potential list structures 3

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

(8)



- 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

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

Carlson, L. and D. Marcu (2001). Discourse tagging reference manual. Available from http://www.isi.edu/˜marcu/discourse/tagging-ref-manual.pdf.

Carlson, L., D. Marcu, and M. E. Okurowski (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith (eds), *Current Directions in Discourse and Dialogue*, 85–112. Dordrecht: Kluwer.

Egg, M. and G. Redeker (2008). Underspecified discourse representation. In A. Benz and P. Kühnlein (eds), *Constraints in Discourse*, 117–138. Amsterdam: Benjamins.

Knott, A., J. Oberlander, M. O'Donnell, and C. Mellish (2001). Beyond elaboration: The interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, and W. Spooren (eds), *Text representation: linguistic and psycholinguistic aspects*, 181–196. Amsterdam: Benjamins.

Lee, A., R. Prasad, A. Joshi, and B. Webber (2008). Departures from tree structures in discourse: Shared arguments in the Penn Discourse Treebank. In *Proceedings of the workshop 'Constraints in Discourse III'*, Potsdam.

Mann, W. and S. Thompson (1988). Rhetorical Structure Theory: Towards a functional theory of text organization. *Text 8*, 243–281.

Marcu, D. (1996). Building up rhetorical structure trees. In *Proceedings of the 13th National Conference on Artificial Intelligence*, Portland, 1069–1074.

Prasad, R., E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, and B. Webber (2006). The Penn Discourse TreeBank 1.0. Annotation Manual. IRCS Technical Report IRCS-06-01, Institute for Research in Cognitive Science, University of Pennsylvania.

Stede, M. (2004). The Potsdam Commentary Corpus. In B. Webber and D. Byron (eds), *ACL 2004 Workshop on Discourse Annotation*, Barcelona, Spain, 96–102. Association for Computational Linguistics.

Taboada, M. and W. Mann (2006). Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies 8*, 423–459.

Webber, B., M. Stone, A. Joshi, and A. Knott (2003). Anaphora and discourse structure. *Computational Linguistics 29*, 545–587.

Wolf, F. and E. Gibson (2005). Representing discourse coherence: a corpus-based study. *Computational Linguistics 31*, 249–287.

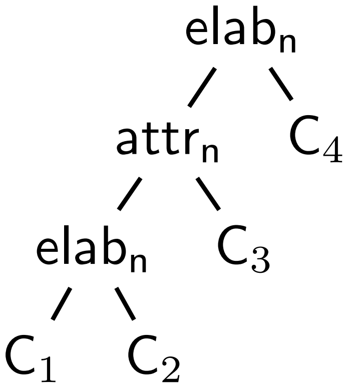
Wolf, F. and E. Gibson (2006). *Coherence in natural language: data stuctures and applications*. Cambridge: MIT Press.

Wolf, F., E. Gibson, A. Fisher, and M. Knight (2005). Discourse Graphbank. Corpus number LDC 2005T08, Linguistic Data Consortium, Philadelphia.

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