How Complex is Discourse Structure?

Markus Egg, Gisela Redeker
Humboldt-Universität Berlin Rijksuniversiteit Groningen
Berlin, Germany Groningen, The Netherlands
E-mail: markus.egg@anglistik.hu-berlin.de, g.redeker@rug.nl

Abstract

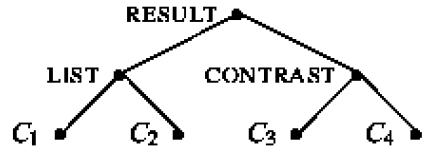
This paper contributes to the question of which degree of complexity is called for in representations of discourse structure. We review recent claims that tree structures do not suffice as a model for discourse structure, with a focus on the work done on the Discourse Graphbank (DGB) of Wolf and Gibson (2005, 2006). We will show that much of the additional complexity in the DGB is not inherent in the data, but due to specific design choices that underlie W&G's annotation. Three kinds of configuration are identified whose DGB analysis violates tree-structure constraints, but for which an analysis in terms of tree structures is possible, viz., crossed dependencies that are eventually based on lexical or referential overlap, multiple-parent structures that could be handled in terms of Marcu's (1996) Nuclearity Principle, and potential list structures, in which whole lists of segments are related to a preceding segment in the same way. We also discuss the recent results which Lee et al. (2008) adduce as evidence for a complexity of discourse structure that cannot be handled in terms of tree structures.

1. Introduction

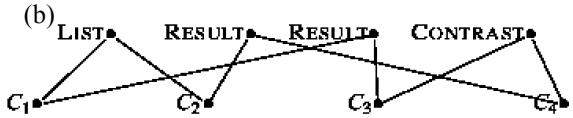
Research on discourse unanimously regards discourse as segmented, and postulates *discourse relations* that combine smaller segments into larger ones, which results in a *discourse structure*. Most discourse structure theories (following Grosz & Sidner, 1986; Polanyi, 1988) assume that discourse structure can be represented as an ordered tree. *N*-ary or binary trees are assumed in annotated discourse corpora, in particular those that implement some version of Rhetorical Structure Theory (RST; Mann & Thompson, 1988; Taboada & Mann, 2006) like the RST Discourse Treebank (Carlson et al., 2002) or the Potsdam Commentary Corpus (Stede 2004). However, this assumption has come under attack as too restricted (Wolf & Gibson, 2005, 2006; Asher, 2008; Danlos, 2008, Lee et al. 2008). In particular, Wolf and Gibson (W&G for short) claim that discourse structure as a rule is much more complex and requires a representation in terms of chain graphs. The difference between the analyses shows up e.g. for (1), whose tree structure is (2a), while W&G's analysis is sketched in (2b).

(1) (C_1) Schools tried to teach students history of science. (C_2) At the same time they tried to teach them how to think logically and inductively. (C_3) Some success has been reached on the first of these aims. (C_4) However, none at all has been reached on the second.

(2) (a)



(b)



Wolf and Gibson base their claims on an annotated corpus of 135 texts from the AP Newswire and Wall Street Journal (source: UPenn TIPSTER), called *Discourse Graphbank* (DGB; Wolf et al., 2005). Non-treeness surfaces in three (interdependent) ways:

First, the *number of relations* is far higher than in a tree-based analysis: There are 9,619 discourse relations in the corpus, 14.4% more than to be expected minimally for a corpus of this size (8,235 segments) if discourse structures are trees (a tree with *n* segments has maximally *n-1* relations). Advocates of tree structures for discourse must be able to account for these surplus relations. Second, the tree-structure constraint of acyclicity is violated by *crossed dependencies,* as non-adjacent segments can be freely related in the chain graph. Wolf and Gibson (2005: 273) report that an average of 12.5% of all relations (median: 10.9%, minimum: 0%, maximum: 44.4% per text) need to be removed to eliminate all crossed dependencies in the DGB. Finally, 41.22% of the segments have *multiple parents* (Wolf & Gibson, 2005: 279), which is also disallowed in tree structures. These phenomena are illustrated by the structures in (2): (2b) uses an additional relation and the individual RESULT relations introduce crossed dependencies and multiple parentship.

In earlier work (Egg & Redeker 2008) we pointed out that the analyses in Wolf and Gibson (2005) have plausible tree-based alternatives. In particular, we showed that much of the additional complexity in W&G's graphs is due to an attempt to integrate into them relations between discourse segments that are eventually due to *cohesive* devices. For instance, in (1), the purported RESULT relations between C_1 and C_3 and C_2 and C_4, respectively, which cannot be captured in terms of a tree structure, can be attributed to the discourse anaphors *the first of these aims* and *the second* in C_2 and C_4, whose respective antecedents are C_1 and C_3. Note that the parallelism of C_1–C_2 and C_3–C_4 remains implicit in both analyses, as the two result relations in W&G's analysis (2b) cannot be restricted to be semantically parallel.

In this paper, we report explorations of crossed dependencies and multiple-parent structures in the DGB,

which are aimed at investigating whether they are really necessary to represent the discourse structures. We will show that much of the additional complexity in the DGB is not inherent in the data, but due to specific *design choices* underlying W&G's annotation.

2. The Analyses

2.1 Crossed Dependencies

Many crossed dependencies in the DGB involve larger distances across the text, which requires careful analysis of the interaction between local and global relations. Consider the first half of wsj_0004, summarized in (3):

(3) (0-1) Money-market yields declined as portfolio managers expect lower interest rates.
(2-4) Declining yields.
(5-16) Lengthening maturities indicate decline of interest rates.
(17-22) Recent rises in short-term interest rates may introduce temporary recovery.

In the DGB, this fragment is analysed as follows:

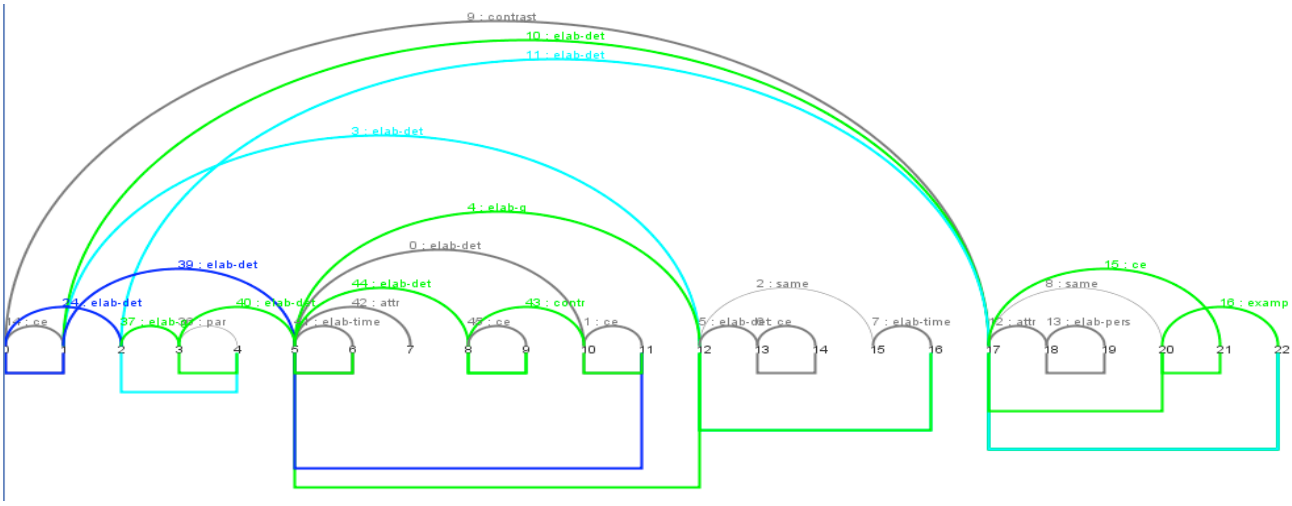


Figure 1: DGB graph for units 0-22 of wsj_0004

Relation number	Span 1		Span 2		Label	Crossing Relations
[0]	10	11	5	6	elab-det	[40],[43]
[1]	11	11	10	10	ce	
[2]	12	12	15	15	same	
[3]	12	16	1	1	elab-det	[11],[24],[39]
[4]	12	16	5	12	elab-g	[11],[40]
[5]	13	14	12	12	elab-det	
[6]	14	14	13	13	ce	
[7]	16	16	15	15	elab-time	
[8]	17	17	20	20	same	
[9]	17	22	0	1	contrast	
[10]	17	22	1	1	elab-det	[24],[39]
[11]	17	22	2	4	elab-det	[3],[4],[24],[39],[40]
[12]	18	19	17	17	attr	
[13]	19	19	18	18	elab-pers	
[14]	1	1	0	0	ce	
[15]	21	21	17	20	ce	[16]
[16]	22	22	20	21	examp	[15]
[24]	2	2	0	1	elab-det	[3],[10],[11],[39]
[36]	3	3	4	4	par	
[37]	3	4	2	2	elab-g	[40]
[39]	5	11	1	1	elab-det	[3],[10],[11],[24],[40]
[40]	5	5	3	3	elab-det	[0],[4],[11],[37],[39],[41],[42],[44]
[41]	6	6	5	5	elab-time	[40]
[42]	7	7	5	6	attr	[40]
[43]	8	9	10	11	contr	[0]
[44]	8	9	5	6	elab-det	[40]
[45]	9	9	8	8	ce	

Table 1: DGB relations and crossed dependencies for units 0-22 of wsj_0004

Table 1 shows 19 crossings for the 27 relations in the fragment. Removing relation [40], which rather implausibly labels (5) "Average maturity of the funds' investments lengthened by a day to 41 days" as an elaboration of (3) "Compound yields assume reinvestment of dividends", reduces this number to 11. Eliminating five more 'long-distance' ELAB-DET relations ([0], [3], [10], [11] and [39]) removes another 10 crossings. The remaining crossing of [15] and [16] disappears once [16] is scoped appropriately: It should relate (22) not to (20-21), but to the whole unit (17-21), in which a quotation (in (17) and (20-21)) is interrupted by an attribution (18-19); the resumption is expressed in the quasi-coherence relation SAME (relation [8]), but ignored in [16]. We will discuss attribution structures in section 2.2.

Most of the problematic relations in the texts we perused are linking elementary units or small segments directly instead of or in addition to linkage at higher levels, and many are of the type ELAB-DET. They often seem motivated solely by lexical or referential overlap. For instance, relation [10] in wsj_0004 links (12-16) to (1), and [39] links (5-11) to (1). The combined unit (0-1) is used in [24], although (2) elaborates only the information in (0) about the decreasing yield, but [10] and [39] single out (1), which introduces interest rates. In a tree-based analysis, the bipartition of the topic sentence (0-1) is paralleled in (2-4) (on yields) and (5-16) (on interest rates), whose conjunction elaborates (0-1).

A strong justification for eliminating these problematic relations lies in the fact that ELAB-DET relations are operating between coherence and cohesion by targeting concepts and not entire discourse segments (see Knott et al 2001). Especially long-distance relations of this type are likely to be inspired by lexical or referential cohesion instead of coherence. In the first 14 texts of the DGB, 36% of all relations, but 69% of long-distance relations (involving a gap of six or more units) are ELAB-DET. This ties in with our earlier observation (Egg & Redeker, 2008) that discourse anaphora like *the first of these aims* in (1) appear to trigger extra relations in the DGB. W&G (2005: 274) report that elaboration relations are involved in 50.5% of all crossed dependencies.

2.2 Multiple-Parent Structures

W&G (2005) illustrate the occurrence of multiple-parent structures with two examples involving parenthetical attribution segments that occur between parts of a quote as C_3 in (4).

(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)

W&G relate the source by ATTRIBUTION to *every* part of the quotation (in addition to the relations between these parts):

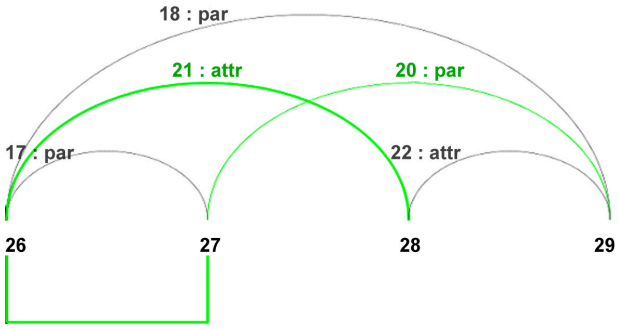
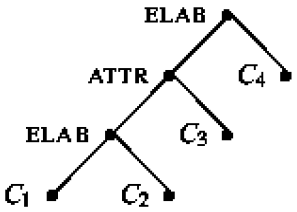


Figure 2: DGB graph for units 26-29 of ap-890101-0003

Egg and Redeker (2008) present a tree-based solution:

(5)



This analysis crucially uses the *nuclearity* principle of Marcu (1996): A relation between a complex segment A and another segment B implies the same relation between the nucleus (central subconstituent) of A and B. Consequently, in (4), 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, which establishes the coherence of C_1-C_2 and C_4. The source is no right boundary of the information, and one need not specify that C_3 indicates the source for C_4, too.[1] (In Redeker & Egg, 2006, we also introduce another tree-based analysis of such cases, involving the extraction of parenthetical attribution phrases.)

We identified 11 quote-medial attribution units with double attribution relations in texts 1-14 of the DGB. Eliminating these relations would account for another 8% of the 138 excess relations for these 14 texts.

For a more general inventory of multiple-parent structures, we identified all spans that appear as the left-hand (i.e. satellite or peripheral) span in more than one relation. In texts 1-14, 235/232 (annotator 1/2) (26%) of all relations are involved in such constructions. Eliminating all but one of the 2-6 relations in each construction would account for 142/139 extra relations, but this may include genuine coherence links.

[1] It has often been noted that the Nuclearity Principle does not hold for all discourse relations. E.g., if a segment S_1 consists of two parts that are linked by a CONSEQUENCE relation, both nucleus and satellite are indispensable, and if S_1 stands in another relation to another segment S_2, this relation need not hold between S_2 and the nucleus of S_1. In our analyses, we haven taken this observation into account.

Elimination would seem safe for ELAB-DET relations, as they probably reflect cohesive linkage. Of the 120/118 ELAB-DET relations in multiple-parent structures, 90/87 could be removed without leaving the span unconnected, accounting for 63-65% of the extra relations.

2.3 Potential List Structures

Another case of multiple attachment (and crossed dependencies) are structures of the type '$A\ B_1\ B_2\ ...\ B_n$' in which all B_i stand in a relation *Rel* to A. E.g., in (6), C_1 is elaborated by [C_2 C_3], C_4, and C_5 (that the first list element is a complex span is incidental.)

(6) (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)

W&G analyse such cases by relating each B_i to A by *Rel* and relating the B_i pairwise (mostly with PARALLEL). The substructure for (6) in their analysis is depicted in Figure 3:

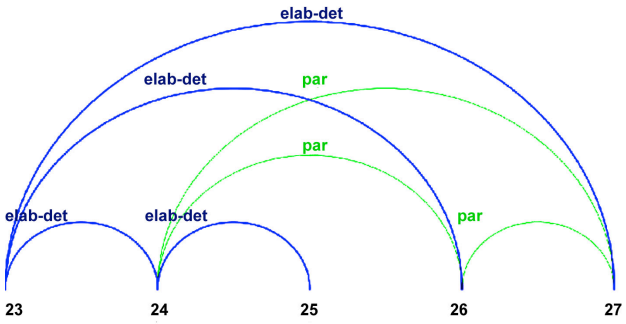
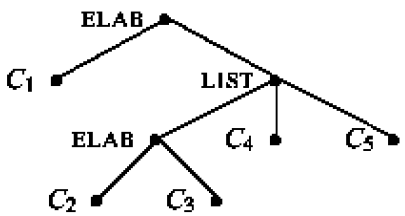


Figure 3: DGB graph for units 23-27 of ap-890101-0002

The intuitions underlying W&G's analysis (segments B_i stand in a relation *Rel* to a segment A and the B_i are comparable or parallel) can also be captured by combining the B_i in a LIST, CONJUNCTION, or SEQUENCE (which each presuppose particular relatedness of the B_i) before relating the whole to A:

(7



W&G's analysis needs $n(n+1)/2$ relations for fragments with n list elements, which the alternative analysis describes with two relations. We identified five such cases, accounting for 15 (10.9%) of the extra relations in texts 1-14.

3. Related Work

In recent work, Lee et al. (2008) present results from their annotation work in the Penn Discourse Treebank (PDTB; Prasad et al., 2008), which they interpret as a motivation to allow for a more expressive representation format for discourse structure. They discuss examples like (8):

(8) (C_1) The London index finished 2.4% under its close of 2233.9 the previous Friday, (C_2) although it recouped some of the sharp losses staged early last week on the back of Wall Street's fall. (C_3) London was weak throughout Friday's trading, however, …

The PDTB analysis relates the subordinate clause in C_2 to both C_1 and C_3 based on the explicit connectives *although* and *however*. Lee et al. (2008) report 349 such instances (4.2% of all cases where a subordinate clause is followed by a sentence with a connective) in the 1-million word corpus. They argue that the occurrence of such cases (in our terminology a type of multiple-parent structure) is evidence that discourse structure is more complex than trees.

Regarding the analysis of (8), we agree that C_3 cannot be directly linked to C_1 by the contrast relation as introduced by *however*. Our analysis of this example *in its discourse context* is in fact quite different. Inspection of the text (wsj_1505) shows a paragraph boundary before C_3. We take the connective *however* to signal a contrast between two multi-sentence paragraphs (e.g. the transition to a new subtopic).

Lee et al. do point out that the PDTB's objective is to annotate individual discourse relations, and not discourse structures. Annotators were presented with one connective at a time and thus could not see whether a specific discourse segment has already previously been selected as the immediate argument of another discourse relation. They had to identify the smallest arguments possible for the discourse relation in question: The *Minimality Principle* in the manual (Prasad et al., 2006) defines such arguments as "minimally required and sufficient for the interpretation of the relation."

Due to this instruction, the PDTB analysis systematically ignores higher-level discourse relations. This is particularly striking in cases where the left-hand argument of a connective is found (often much) earlier in the text and is non-adjacent to the units the connective occurs with. In all such cases we perused, the left-hand argument in question would be the nucleus of a longer stretch of discourse that would span all intervening text up to the connective.

For the purposes of the PDTB, which is directed at individual connectives and relations, this need not be problematic. It has often been noted that it is a typical characteristic of satellites in a discourse structure that they can be left out in a text without resulting in a non-coherent text (see, e.g., Marcu's (1996) Nuclearity Principle). But we do question the interpretation of

treeness violations resulting from juxtaposing separate, minimal PDTB annotations as evidence against the sufficiency of trees to model discourse structure.

4. Conclusion and Outlook

Our results raise serious doubts about W&G's evidence for the claim that trees are not descriptively adequate data structures for representing discourse structure. Many of the additional relations structures in the DGB that cannot be captured in terms of tree structures can be shown to arise from design decisions and not from empirical necessity.

Our next step is a more systematic analysis of relations in the DGB that appear to be based on anaphoric or sense relations between lexemes instead of capturing coherence structures. Careful manual evaluation is necessary to distinguish relations established by cohesive means alone from those in which cohesion accompanies an independently established discourse relation. We are using comparisons with RST analyses to identify cases of low-level cohesion-based linkage in the DGB where a tree-based analysis yields a plausible hierarchical structure.

5. Acknowledgements

Part of this work has been supported by grant 360-70-282 of the Netherlands Organization for Scientific Research (NWO) as part of the NWO-funded program *Modelling discourse organization* (http://www.let.rug.nl/mto/). We wish to thank Marco Trevisan for his contributions to the analyses presented in this paper and three anonymous reviewers for their helpful comments on an earlier version of this paper.

References

Asher, N. (2008). Troubles on the right frontier. In P. Kühnlein & A. Benz (eds), *Constraints in discourse.* Amsterdam: Benjamins, pp.29–52.

Carlson, L., D. Marcu, M. E. Okurowski (2002). *RST Discourse Treebank*. Linguistic Data Consortium, Philadelphia.

Danlos, L. (2008). Strong generative capacity of RST, SDRT, and discourse dependency DAGs. In P. Kühnlein & A. Benz (eds), *Constraints in discourse.* Amsterdam: Benjamins, pp. 69–95.

Egg, M., G. Redeker (2008). Underspecified discourse representation. In Anton Benz & Peter Kühnlein (eds), *Constraints in Discourse*, Amsterdam: Benjamins, pp. 117–138.

Grosz, B. & C. Sidner (1986). Attention, intention, and the structure of discourse. *Computational Linguistics* 12, 175–204.

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

Lee, A., R. Prasad, A. Joshi, B. Webber (2008). Departures from Tree Structures in Discourse: Shared Arguments in the Penn Discourse Treebank. In *Proceedings of the Constraints in Discourse III Workshop, Potsdam, Germany, July-August 2008.*

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

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

Polanyi, L. (1988). A formal model of discourse structure. *Journal of Pragmatics* 12, pp. 601–638.

Prasad, R., A. Lee, N. Dinesh, E. Miltsakaki, G. Campion, A. Joshi, B. Webber (2008). *Penn Discourse Treebank Version 2.0.* Linguistic Data Consortium, Philadelphia.

Prasad, R., E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, 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.

Redeker, G., M. Egg (2006). Says who? On the treatment of speech attributions in discourse structure. In C. Sidner, J. Harpur, A. Benz, & P. Kühnlein (eds), *Proceedings of the Workshop Constraints in Discourse 2006.* Maynooth: National University of Ireland, pp. 140–146.

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

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

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

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

Wolf, F., E. Gibson, A. Fisher, M. Knight (2005). *Discourse Graphbank.* Linguistic Data Consortium, Philadelphia.