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

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) *(C1)* Schools tried to teach students history of science. *(C2)* At the same time they tried to teach them how to think logically and inductively. *(C3)* Some success has been reached on the first of these aims. *(C4)* However, none at all has been reached on the second.

(2) *(a)*

RESULT

LIST

CONTRAST

*C1* *C2* *C3* *C4*

(b) LIST RESULT RESULT CONTRAST

*C1* *C2* *C3* *C4* *C5*

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 C1 and C3 and C2 and C4, 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 C3 and C4, whose respective antecedents are C1 and C3. Note that the parallelism of C1–C2 and C3–C4 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](image)

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

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 $C_1$ 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](image)

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 $C_1$ and $C_2$ and the nucleus of $C_1-C_2$ (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 not a right boundary of the information, and one need not specify that $C_4$ 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 $S_x$ consists of two parts that are linked by a CONSEQUENCE relation, both nucleus and satellite are indispensable, and if $S_i$ stands in another relation to another segment $S_y$, this relation need not hold between $S_x$ and the nucleus of $S_i$. In our analyses, we haven’t taken this observation into account.
W&G analyse such cases by relating each $B_i$ to $A$ by $\text{Rel}$ and relating the $B_i$ pairwise (mostly with $\text{PARALLEL}$). The substructure for (6) in their analysis is depicted in Figure 3:

![DGB graph for units 23-27 of ap-890101-0002](image)

The intuitions underlying W&G’s analysis (segments $B_i$ stand in a relation $\text{Rel}$ to a segment $A$ and the $B_i$ are comparable or parallel) can also be captured by combining the $B_i$ in a $\text{LIST}$, $\text{CONJUNCTION}$, or $\text{SEQUENCE}$ (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_3) The London index finished 2.4% under its close of 2233.9 the previous Friday, (C_3) 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 $\text{although}$ and $\text{however}$. 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 $C_1$ cannot be directly linked to $C_2$ by the contrast relation as introduced by $\text{however}$. 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 $\text{however}$ 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 $\text{Minimality Principle}$ 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


