Same and *Elaboration* Relations in the Discourse Graphbank

Irina Borisova
University of Groningen,
Groningen, The Netherlands
Saarland University,
Saarbrücken, Germany
borisova.ira@gmail.com

Gisela Redeker
University of Groningen,
Groningen, The Netherlands
g.redeker@rug.nl

Abstract

This study investigates the use of *Same* – a relation that connects the parts of a discontinuous discourse segment – in the Discourse Graphbank (Wolf et al., 2004). Our analysis reveals systematic deviations from the definition of the *Same* relation and a substantial number of confusions between *Same* and *Elaboration* relations. We discuss some methodological and theoretical implications of these findings.

1 Introduction

Coherence relations and their composition (usually assumed to be strictly hierarchical, i.e., treelike) form the core of most corpus-linguistic and computational work on discourse structure (see Taboada & Mann 2006 for an overview). The assumption that discourse structure can be modeled as a tree has recently come under attack e.g. in Wolf & Gibson (2003, 2006; henceforth WG). Based on the *Discourse Graphbank* (Wolf et al 2004; henceforth DG), a manually annotated corpus of 135 newspaper and newswire texts, WG claim that less constrained graph structures are needed that allow for crossed dependencies (i.e. structures in which discourse units ABCD (not necessarily adjacent) have relations AC and BD) and multiple-parent structures (where a unit enters more than one coherence relation and is thus dominated by more than one node).[1]

Among the 11 types of relations distinguished in DG, the *Elaboration* relation, where two asymmetrically related discourse units are "centered around a common event of entity"(Wolf et al 2003: 12), stands out by its heavy involvement in these violations of tree structure constraints. *Elaboration* relations are involved in 50.52% of all crossed dependency structures and in 45.83% of multiple-parent structures. These high percentages are in part due to the high overall frequency of *Elaboration* relations (37.97% of all relations), but clearly exceed that base rate. Elsewhere, *Elaboration* relations, esp. those where the elaborandum is an entity and not a whole proposition, have been criticized as belonging more to referential coherence than to relational coherence (Knott el at 2001). In this study, we show that WG's (somewhat idiosyncratic) definition of the *Elaboration* relation seems to lead to confusion with the 'pseudo-relation' *Same.*

The 'pseudo-relation' *Same-Unit* was introduced by Marcu (Carlson & Marcu 2001) to deal with discontinuous discourse units in the RST Discourse Treebank (Carlson, Marcu & Okurowski 2002). *Same-Unit* (re)connects the parts of a discourse unit that is disrupted by embedded material. In the tree representation, the intervening material is attached to one of the constituent units of the *Same-Unit* relation (Carlson & Marcu 2001:23-26). In DG, this relation is called *Same* and accounts for 17.21% of all relations; only *Elaboration* and *Similarity* are more frequent.[2] As DG allows multiple attachments, *Same* should be expected to be regularly associated with multiple-parent structures, and it is: the percentage of *Same* relations is higher in multiple-parent structures than overall, and the reduction of multiple-

[1] The validity of this claim is contested in Egg & Redeker (2010).

[2] Note that a *Same-Unit* relation is not needed in 'classic' RST, where parenthetical segments are extracted and placed after the segment within which they occur (Redeker & Egg 2006).

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 63–66,
The University of Tokyo, September 24-25, 2010.

parent structures when *Same* relations are removed from the DG is second only to *Elaboration* (Wolf & Gibson 2003:280-282).

Our explorations of *Same* relations in DG revealed a substantial number of cases that do not seem to fit WG's definition of this relation, most notably confusions with *Elaboration* relations and a surprising number of cases where there is no intervening segment to be bridged by the *Same* relation. In this paper, we will present these findings and discuss some consequences for discourse segmentation and the annotation of coherence relations.

2 *Same* relations in DG

The DG coding manual (Wolf et al 2003:15) stipulates as the only condition for a *Same* relation that a discourse segment must have "intervening material". The example in the manual tacitly fits the much more restrictive definition given in (Wolf & Gibson 2003:255) and in (Wolf & Gibson 2006:28):

> "A same relation holds if a subject NP is separated from its predicate by an intervening discourse segment".

Among the 534 *Same* relations in DG,[3] we have identified 128 cases (23.98%) where this definition does not seem to apply. Sixty-four of these cases also do not satisfy the broader definition in the coding manual (see 2.3).

2.1 *Same* or *Elaboration*?

In 35 cases, the *Same* relation is applied to constructions that are elsewhere labeled *Elaborations*. Consider the parallel examples (1) and (2):

```
(1) [42]–[44] elab-loc
[42] There, [43] she said,
[44] robots perform specific
tasks in "islands of
automation," (Text 1)


(2) [32]–[34] same
[32] In the factory of the
future, [33] according to the
university's model, [34]
human chatter will be
replaced by the click-clack
of machines. (Text 1)
```

In these examples, [42] and [32] each specify a location for the state of affairs expressed in the second constituent of the relation, [44] and [34] respectively. Note that [32] is not a subject NP and example (2) thus violates the restricted variant of the *Same* relation definition. Interestingly, examples (1) and (2) differ with respect to the involvement in crossed dependencies and multiple-parent structures. As expected from an elaborating segment, [42] does not participate in any other relations; the three other relations [44] participates in do not include [42]. By contrast, [32] is attached to the intervening segment and in eight other relations in which not [34] by itself, but the combined segment [32]–[34] participates.

In other examples, a general difference between these *Same* and *Elaboration* examples lies in the attachment of the intervening segment: in the *Same* cases, the intervening segment might be attached to the preceding discourse segment, and in the *Elaboration* cases to the following segment.

The confusion between the symmetric *Same* relation (both segments have in principle equal status) and the asymmetric *Elaboration* relation (combining an elaborandum with a less central elaborating segment) might have been caused by WG's definition, which stipulates that the segments be "centered around a common event or entity" (Wolf et al 2003: 12) and thus does not reflect the asymmetry of the *Elaboration* relation.

2.2 Violations of definitional constraints

There are other cases, besides those discussed in 2.1, where the formal requirement of the restrictive definition is not met. In 20 cases, the *Same* relations joins coordinated or disjoint NP's as in example (3):

```
(3) [13]–[16] same
[13] Mrs. Price's husband,
[14] Everett Price, [15] 63,
[16] and their daughters,
(Text 2)
```

In 12 cases, *Same* is used to relate a discourse connective to its host clause as in (4):

```
(4) [4]–[6] same
[4] However, [5] after two
meetings with the Soviets,
[6] a State Department
spokesman said that (Text 8)
```

[3] We have arbitrarily chosen to use the data for annotator 1. The two annotators agreed on segmentation and annotation in 98% of the cases.

Presumably the annotators were using the less restrictive definition in the coding manual. This explanation cannot account for the last category of problematic cases we now turn to.

2.3 Spurious *Same* relations

We found 64 cases in DG where *Same* is assigned to two adjacent discourse segments, thus violating the essential criterion of "intervening material". Such 'spurious' *Same* relations occur with various constructions including the following:

- Complement clauses

(5) [61] The administration should now state [62] that (Text 123, wsj_0655)

- Infinitive clauses

(6) [79] Banco Exterior was one of the last banks [80] to create a brokerage house (Text 122, wsj_0616)

- Conditional clauses

(7) [35] And important U.S. lawmakers must decide at the end of November [36] if the Contras are to receive the rest of the $49 million in so-called humanitarian assistance under a bipartisan agreement (Text 123, wsj_0655).

- Gerund postmodifier phrases

(8) [2] Lawmakers haven't publicly raised the possibility [3] of renewing military aid to the Contras, (Text 123, wsj_0655).

- Temporal "as"-clauses

(9) [31] it came [32] as Nicaragua is under special international scrutiny in anticipation of its planned February elections. (Text 123, wsj_0655)

The 64 spurious *Same* relations are concentrated in only 20 of the 135 texts. Fifty-one of those cases occur in ten texts that were also used in the RST Discourse Treebank. This gives us the interesting opportunity to compare the DG and RST Treebank analyses for these 51 cases. As Table 1 shows, only two of them are labeled *Same-Unit* in the RST Treebank, while 26 (51%) are *Elaboration* relations.

Relations	Frequencies	Percent
Elaboration	26	51.0 %
Attribution	13	25.5 %
Same-Unit	2	3.9 %
Other	10	19.6 %
Total	51	100 %

Table 1: Spurious *Same* relations in DG and relations assigned in the RST Treebank

It is instructive to look at the subtype of *Elaboration* assigned to these cases, which most commonly is the relation *Elaboration-object-attribute-e*. It applies to clausal modifiers, usually postmodifiers of a noun phrase, that express an intrinsic quality of an object. Carlson & Marcu (2001:55) illustrate this relation with the following example:

(10) [Allied Capital is a closed-end management investment company][that will operate as a business development concern.] (wsj_0607)

The constructions with spurious *Same* relations in DG thus often involve restrictive modification, implying a very close tie between the segments involved, possibly prompting the annotators to as it were undo the segmentation.

3 Segmentation rules

Any annotation of discourse relations requires rules for segmenting the text into elementary discourse units. DG follows Carlson & Marcu (2003) in assuming clauses, modifiers and attributions as discourse segments (DSs), but adds some "refinements" (Wolf et al., 2003:8) that may be responsible for some of the problematic cases discussed in section 2.[4] In particular, two of the additional stipulations refer to "elaborations":

[4] A different account of the segmentation is given in (Wolf & Gibson 2006), but the annotation in DG is presumably based on the 2003 manual.

> "Elaborations [..] are separate DSs*: [Mr. Jones,][spokesman for IBM,] [said…]"* (Wolf et al., 2003:8)
>
> "Time-, space-, personal- or detail-elaborations are treated as DSs" (Wolf et al., 2003:9).

This might simply be an unfortunate equivocation, but still is likely to confuse annotators by confounding the segmentation and relation annotation tasks.

4 Conclusions

Our analysis of the *Same* relation in DG has shown systematic deviations from the definition of this (pseudo-)relation and a substantial number of confusions between *Same* and *Elaboration*, both in cases where *Same* cannot apply, as there is no intervening segment, and in cases where both might apply, but parsimony would demand to treat parallel cases equally. Some of the problematic cases may have been caused by the use of relational terminology ("elaboration") in two of the segmentation rules. The problems are not just methodological, though, but may raise questions about the conceptual status of *Elaboration* relations.

The confusion of a bone fide coherence relation with a purely technical construction that serves to recombine the parts of an interrupted segment must be worrisome. More specifically, the comparison with the annotation in the RST Discourse Treebank reveals that many of the 'spurious' *Same* relations in DG are analyzed as *Elaboration-object-attribute-e* relations in the RST Treebank. This is exactly the subcategory of *Elaboration* relations that most clearly operate on the level of entities instead of propositions, and thus arguably might not be proper discourse relations (Knott et al. 2001). This holds a fortiori as Carlson & Marcu's (2001) definition of the *Elaboration-object-attribute-e* relation requires a restrictive modifier construction. The increasing availability of corpora annotated for discourse structure will facilitate the further investigation of these questions.

Acknowledgements

This research was partially supported by a travel grant from the Erasmus Mundus Masters Programme in Language and Communication Technologies to Borisova and by grant 360-70-282 of the Netherlands Organization for Scientific Research (NWO) to Redeker. We would like to thank Robin Cooper and three anonymous reviewers for their very useful comments.

References

Lynn Carlson and Daniel Marcu. 2001. *Discourse Tagging Reference Manual.*

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. *RST Discourse Treebank.* Linguistic Data Consortium, Philadelphia.

Markus Egg and Gisela Redeker. 2010. How complex is discourse? *Proceedings of LREC 2010.*

Alistair Knott, Jon Oberlander, Michael O'Donnell, and Chris Mellish. 2001. Beyond *Elaboration*: The interaction of relations and focus in coherent text. In J. Schilperoord T. Sanders and W. Spooren, editors, *Text Representation: Linguistic and Psycholinguistic Aspects*, pp 181–196. Benjamins.

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

Maite Taboada and William C. Mann. 2006. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies, 8* (3), 423–459.

Florian Wolf, Edward Gibson, Amy Fisher, Meredith Knight. 2003. *A Procedure for Collecting a Database of Texts Annotated with Coherence Relations.* Database documentation.

Florian Wolf and Edward Gibson. 2003. Representing discourse coherence: A corpus-based study. *Computational Linguistics, 31*(2):249–287.

Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. 2004. *Discourse Graphbank.* Linguistic Data Consortium, Philadelphia.

Florian Wolf and Edward Gibson. 2006. *Coherence in Natural Language.* MIT Press, Cambridge, MA.