A Study of Parentheticals in Discourse Corpora — Implications for NLG Systems

Eva Banik[†] **and Alan Lee**[‡]

[†]Department of Computing, The Open University, Milton Keynes, UK
[‡]Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA, USA
e.banik@open.ac.uk, aleewk@babel.ling.upenn.edu

Abstract

This paper presents a corpus study of parenthetical constructions in two different corpora: the Penn Discourse Treebank (PDTB, (PDTB-Group, 2008)) and the RST Discourse Treebank (Carlson et al., 2001). The motivation for the study is to gain a better understanding of the rhetorical properties of parentheticals in order to enable a natural language generation system to produce parentheticals as part of a rhetorically well-formed output. We argue that there is a correlation between syntactic and rhetorical types of parentheticals and establish two main categories: ELABORATION/EXPANSION-type NP-modifier parentheticals and NON-ELABORATION/EXPANSION-type VP- or S-modifier parentheticals. We show several strategies for extracting these from the two corpora and discuss how the seemingly contradictory results obtained can be reconciled in light of the rhetorical and syntactic properties of parentheticals as well as the decisions taken in the annotation guidelines.

1. Definition

Parentheticals are constructions that typically occur embedded in the middle of a clause. They are not part of the main predicate-argument structure of the sentence and are marked by special punctuation (e.g. parentheses, dashes, commas) in written texts, or by special intonation in speech. Syntactically, parentheticals can be realized by many different constructions. Some examples of different types of parentheticals are given in (1): appositive relative clause (1a), non-restrictive relative clause (1b), participial clause (1c), subordinate clause (1d) . Throughout this paper we will show parenthetical constructions in square brackets.

(1) a The new goal of the Voting Rights Act [– more minorities in political office –] is laudable. (wsj1137)

b GE, [which vehemently denies the government's allegations,] denounced Mr. Greenfield's suit. (wsj0617)

c But most businesses in the Bay area, [including Silicon Valley,] weren't greatly affected. (wsj1930)

d So far, [instead of teaming up,] GE Capital staffers and Kidder investment bankers have bickered. (wsj0604)

A common characteristics of parentheticals is that they express information that is not central to the meaning of the overall message conveyed by a text or spoken utterance and since they are specifically marked by punctuation or intonation, they allow the reader to distinguish between more and less important parts of the message. By structuring information this way, parentheticals make it easier for readers to decode the message conveyed by a text. Consider for example the following message that has been expressed by two different texts: one without parentheticals (2a) and one that contains two parentheticals (2b).

(2) a Eprex is used by dialysis patients who are anaemic. Prepulsid is a gastro-intestinal drug. Eprex and Prepulsid did well overseas.

b Eprex, [used by dialysis patients who are anaemic,] and Prepulsid, [a gastro-intestinal drug,] did well overseas. (wsj1156)

2. Background and Motivation

Parentheticals have been much studied in linguistics (see (Dehe and Kavalova, 2007), (Burton-Roberts, 2005) for a recent overview) but so far they have received less attention in computational linguistics. Only a few studies have attempted a computational analysis of parentheticals, the most recent ones being (Bonami and Godard, 2007) who give an underspecified semantics account of evaluative adverbs in French and (Siddharthan, 2002) who develops a statistical tool for summarisation that separates parentheticals from the sentence they are embedded in. Both of these studies are limited in their scope as they focus on a very specific type of parentheticals.

From the perspective of natural language generation (NLG), as far as we know, nobody has attempted to give a principled account of parentheticals, even though these constructions contribute to the easy readability of generated texts, and therefore could significantly enhance the performance of NLG systems (Scott and Souza, 1990).

Since the input to most NLG systems is a text plan expressed in some variant of Rhetorical Structure Theory (Mann and Thompson, 1987), the first step towards generating parentheticals is to understand which parts of the input text plan can be expressed as parenthetical constructions.

The purpose of the present study is to take this first step by examining the rhetorical context of parentheticals. We describe a corpus study on two differently annotated discourse treebanks: the RST Discourse Treebank (RST, (Carlson et al., 2001)) and the Penn Discourse Treebank (PDTB, (PDTB-Group, 2008)). We show how parentheticals can be extracted from these treebanks and identify the discourse

relations whose arguments are realized as parentheticals. We argue that the type of rhetorical or discourse relation that holds between parenthetical and its host correlates with the syntactic construction used for the parenthetical. We distinguish between two main types of parentheticals: i) ELABORATION/EXPANSION-type nominal modifiers which express a relation between an object and a proposition and ii) NON-ELABORATION/EXPANSION-type VP- or S-modifiers which express a relation between two propositions. We show how the seemingly contradictory findings from the two corpora can be reconciled in light of these two types and the different perspective on discourse adopted in each corpus.

The next two sections of the paper discuss the strategies we used to extract parentheticals from our chosen corpora. Section 5. discusses our findings and shows how the results of the study can be used to inform a natural language generation system.

3. Corpus Study on RST Treebank

The RST Discourse Treebank (Carlson et al., 2001) consists of 385 Wall Street Journal texts from the Penn Treebank, segmented into Elementary Discourse Units (EDUs) and annotated with rhetorical relations (78 different relations in total).

The corpus annotation manual defines embedded units as EDUs which break up another legitimate EDU or modify a portion of an EDU only, not the entire EDU. The corpus defines a pseudo-relation called Same Unit, which is used as a device for linking two discontinuous text fragments that are really a single EDU, but are broken up by an embedded unit. Same Unit is a multinuclear relation, where one of the nuclei corresponds to the parenthetical and the part of the host sentence that the parenthetical is related to (e.g. head noun of a relative clause). The other nucleus either simply contains the rest of the sentence or it can be similarly complex if there is more than one parenthetical in the sentence. To illustrate, Figure 1. shows an example of a sentence annotated with the Same Unit relation, where Nucleus 1 contains the subject NP broken up into a head noun and a complex relative clause, the latter being the parenthetical.

Same Unit annotations identify EDUs that are embedded linearly or syntactically within another EDU, including parentheticals that are not syntactically related to the sentence but are separated by punctuation.

We have extracted all the examples of Same Unit annotations from the corpus, 1117 cases in total and counted the number and type of rhetorical relations that occur within a complex nucleus.[1] There were 1401 complex nuclei in total, of which we have excluded 630 cases that contained ELABORATION-OBJECT-ATTRIBUTE annotations.

The reason for this is that by definition, in the corpus ELABORATION-OBJECT-ATTRIBUTE relations are assigned to restrictive relative clauses, which from a natural language generation system's perspective, are generated as referring expressions. We therefore don't consider ELABORATION-OBJECT-ATTRIBUTE annotations as parentheticals.

[1]For complex cases like the example in Figure 1 we only counted the top relation.

within SameUnit		Total in corpus		
331	42.93%	3510	20.64%	elab-add
128	16.60%	2520	14.82%	attribution
58	7.52%	558	3.28%	circumstance
35	4.54%	463	2.72%	purpose
22	2.85%	97	0.57%	restatement
20	2.59%	181	1.06%	condition
19	2.46%	266	1.56%	example
18	2.33%	337	1.98%	antithesis
14	1.82%	110	0.65%	elab-set-member
13	1.69%	223	1.31%	concession
11	1.43%	368	2.16%	elab-gen-spec
102	13.23%	8371	49.23%	Other
771	100.00%	17004	100.00%	

Table 1: Most frequent relations within Same Unit

Table 1 shows the frequency of relations that occurred more than ten times inside the Same Unit relation. We didn't find any difference between embedded relations that occur in the first vs. second nucleus of Same Unit, and the table presents the summary of all relations that we found in one of the nuclei.[2] The first column shows the number and frequency of rhetorical relations within Same Unit, and the second column gives the overall number and frequency of relations in the whole corpus. As can be seen by comparing the numbers in the two columns, all the relations in the table occur more frequently than usual within Same Unit. The most common relations expressed by parentheticals are ELABORATION-TYPE relations, which occur 51.49% of the time, with ELABORATION-ADDITIONAL accounting for 42.93% of all cases. The rest of the cases are CIRCUMSTANCE, PURPOSE, CONDITION, ANTITHESIS or CONCESSION relations which altogether occur 35.27% of the time.

In order to generate parentheticals, a system needs to know not only which rhetorical relations can be realized as parentheticals, but also which of the available syntactic constructions to choose to realize it. To determine which syntactic constructions co-occur most frequently with parenthetical rhetorical relations, we have looked at 640 examples of Same Unit annotations and noted the syntactic type of the embedded parenthetical. The summary of our typology is illustrated in Table 2.

As the table shows, although each rhetorical relation can be realized by several different syntactic types, there are one or two types for each relation that are more frequent than others. These are shown in the table in boldface.

Based on which syntactic construction is most frequently used, we can distinguish two types of rhetorical relations. ELABORATION-TYPE relations (ELABORATION-ADDITIONAL, ELABORATION-SET-MEMBER, ELABORATION-GENERAL-SPECIFIC, RESTATEMENT and EXAMPLE) are most frequently expressed by NP-modifiers (non-restrictive relative clauses, NPs and participial clauses) which appear immediately following

[2]Most examples of Same Unit had two nuclei. We have found 9 examples with 3 nuclei and 3 cases of 4 nuclei, but these were excluded from the study.

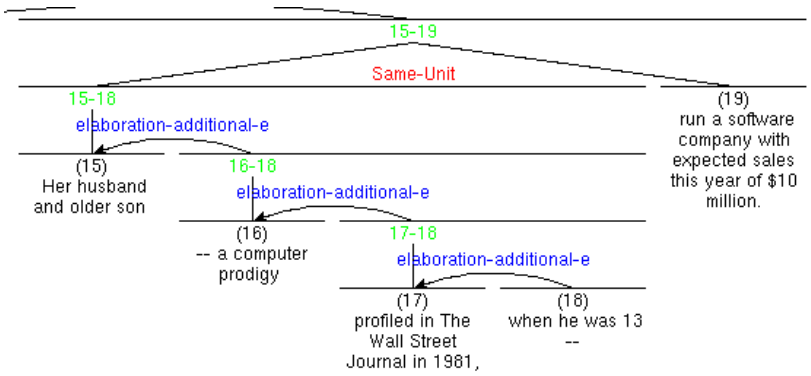

Figure 1: Embedded discourse unit in Nucleus 1 of a Same Unit relation

the entity which forms the nucleus, most of the time an NP. CIRCUMSTANCE, PURPOSE, ANTITHESIS or CONCESSION relations on the other hand are mostly expressed by VP- or S-modifiers (cue + S subordinate clauses, to-infinitives, and in the case of ATTRIBUTION, NP + V sequences).
According to the annotation manual, EDUs that are expressed by certain syntactic constructions such as relative clauses, nominal postmodifiers with a non-finite clause and appositives by definition always play the role of the Satellite and enter into "embedded" rhetorical relations, marked by a "-e" following the name of the relation.
To illustrate, (3) shows an example of a regular CONCESSION relation, and (4) an embedded CONCESSION-E relation.

(3) Though couriers work as many as 60 hours a week during the autumn rush, they leave early during slack times while still being assured of a minimum paycheck.(CONCESSION)

(4) All citizens, [regardless of race], must feel represented. (CONCESSION-E)

Overall in the corpus 20% of all discourse relations are annotated as "embedded". The most frequently occurring of these relations is ELABORATION-OBJECT-ATTRIBUTE, which appears in a '-e' form 97.37% of the time. This is a consequence of the definition of ELABORATION-OBJECT-ATTRIBUTE, which is by definition assigned to restrictive relative clauses. Indeed, many of the "non-embedded" examples of ELABORATION-OBJECT-ATTRIBUTE seem to be due to annotation errors. Again, we have excluded ELABORATION-OBJECT-ATTRIBUTE-E relations from the set of '-e' relations considered. Table 3. gives the statistics for occurrences of '-e' relations in the corpus, showing the frequency of embedding for relations that occur more than ten times in an embedded form in the corpus and excluding ELABORATION-OBJECT-ATTRIBUTE-E.
Table 3 agrees with our study of Same Unit relations in that the most frequently embedded relations are subtypes of elaboration. But these types account for 870 cases, or 73.54% of all '-e' relations which is far more than what we found embedded within Same Unit (51.49%). The rest of the relations in Table 3 correspond to the category of VP- or S-modifier relations identified above, accounting for 128 or 10.8% of '-e' relations, and ATTRIBUTION (8.6%).

regular		'-e'		Total	rhetorical relation
55	56.70%	42	43.30%	97	restatement
69	62.73%	41	37.27%	110	elab-set-member
2820	80.34%	690	19.66%	3510	elab-add
51	80.95%	12	19.05%	63	definition
223	83.83%	43	16.17%	266	example
326	88.59%	42	11.41%	368	elab-gen-spec
422	91.14%	41	8.86%	463	purpose
511	91.58%	47	8.42%	558	circumstance
166	91.71%	15	8.29%	181	condition
212	95.07%	11	4.93%	223	concession
323	95.85%	14	4.15%	337	antithesis
2367	95.87%	102	4.13%	2469	attribution
5998	98.64%	83	1.36%	6081	other
13543	91.97%	1183	8.03%	14726	total in corpus

Table 3: Relations most frequently occurring in a '-e' form

The numbers in Table 3 give us an idea about which relations occur frequently in a syntactically embedded form, and confirms that the set of relations that we found within Same Unit annotations are indeed the most frequently embedded ones. However, there is a significant difference in the distribution of the individual relations and it seems that '-e' annotations are not a good indicator for parenthetical status since we found both '-e' and "regular" rhetorical relations within Same Unit. For the purposes of this corpus study we decided therefore to ignore the distinction between '-e' and regular occurrences of rhetorical relations and only consider a relation parenthetical when it appears within a Same Unit annotation.

4. Corpus Study on PDTB

The Penn Discourse Treebank (PDTB) takes a different perspective on discourse annotation. The corpus considers discourse connectives to be predicates that express binary discourse relations and focuses on identifying a chosen set of over a hundred discourse connectives and their arguments, rather than attempting to annotate all relations expressed in a given text. The PDTB also annotates the argument struc-

		Elab-add	Example	Elab-gen-spec	Restatement	Elab-set-mem	Attribution	Condition	Antithesis	Concession	Circumstance	Purpose	
NP-modifiers	relative clause	**143**		2		2							147
	participial clause	**96**	4			1	1				11	4	117
	NP	**34**		**8**	**22**								64
	NP-coord					6							6
	cue + NP	5	1						2	3	2		13
	Adj + cue	2											2
	number	2											2
	including + NP		**13**			5							18
VP- or S- modifiers	to-infinitive	4										**30**	34
	NP + V						**106**						106
	cue + S	5						**20**	**14**	**9**	**29**		77
	PP	11									9	1	21
	S	7	1	1									9
	according to NP						7						7
	V + NP						6						6
	as + S						4						4
	Adv + number	1									1		2
	cue + Adj										2		2
	cue + participial								2				2
	cue + V						1						1
		310	19	11	22	14	125	20	18	12	54	35	640

Table 2: Syntactic types of parentheticals in Same Unit annotations

ture and semantics of some implicit discourse relations and entity relations (similar to entity-based coherence in (Knott et al., 2001)).

The two arguments of a discourse connective are labelled Arg1 and Arg2 in the corpus, where Arg2 is the argument that is syntactically bound to the discourse connective, and Arg1 is the other argument. In the PDTB all connectives have exactly two arguments and there are no constraints on the relative order of these arguments.

One way to identify parenthetical constructions in this corpus is to find annotations where one argument of a discourse connective occurs linearly embedded within the other argument. There are two logical possibilities for this: either Arg1 is embedded within Arg2 or Arg2 is embedded in Arg1. We will call the first type, where Arg1 forms the parenthetical, **Arg1-parentheticals**. This type is illustrated in (5a). The second type, where Arg2 is parenthetical, is shown in (5b). We'll call this type **Arg2-parentheticals**.

(5) a [...] [$_{ARG2}$ late Tuesday the Chinese government, [$_{ARG1}$ which often buys U.S. grains in quantity,] turned **instead** to Britain to buy 500,000 metric tons of wheat] (wsj0155)

b [...] [$_{ARG1}$ pollination, **while** [$_{ARG2}$ easy in corn because the carrier is wind,] is more complex and involves insects as carriers in crops such as cotton] (wsj0209)

In Arg1-parentheticals the discourse connective appears in the host sentence, whereas in Arg2-parentheticals it appears within the parenthetical.

In total there are 219 cases of "embedded-argument" parentheticals in the corpus, of which 207 are examples of Arg2-parentheticals and only 12 are Arg1-parentheticals. The distribution of these two types is illustrated in table 4. The rows of the table show the number of parentheticals found for four different semantic type of connectives: TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION, i.e. the number of cases where a discourse relation of a given type holds between the parenthetical and its host. The striking difference between the numbers in the two columns of table 4 shows that when one of the arguments of an explicit discourse relation is expressed as a parenthetical, in most cases the embedded string contains the explicit connective. The connectives that appear within the parenthetical are mostly structural connectives (mainly subordinating conjunctions and their modified forms, e.g., after, although, as, because, before, if, once, since, though, until, when, while), whereas all the connectives that appear in the host sentence are adverbs (e.g., also, instead, later, nevertheless, nonetheless, previously, still, thereafter).

The distribution of parentheticals across types of discourse relations in table 4 shows that there is no restriction on what relations allow their Arg2 to be expressed as a parenthetical. However, we find, curiously, only few examples of expansion-type parentheticals despite the fact that this is the most common semantic category among the explicit connectives in the entire corpus. Also, as we've seen in the previous section, the majority of parentheticals in the same

Relation	Arg2-parentheticals	Arg1-parentheticals	distribution in corpus
TEMPORAL	101 (48.8%)	2	3434 (18.6%)
CONTINGENCY	53 (25.6%)	0	3286 (17.8%)
COMPARISON	38 (18.3%)	5	5490 (29.7%)
EXPANSION	15 (7.2%)	5	6239 (33.8%)
TOTAL:	207	12	18484

Table 4: Relations between "embedded-argument" parentheticals and their hosts

text according to the annotations of the RST treebank were in fact elaboration- (or expansion) type. We will discuss the reason for this discrepancy in more detail in section 5.

Nevertheless, it should be noted that a lot of expansion-type relations can be inferred from another type of annotation in the PDTB: supplementary material annotations. These are extensions to a discourse argument that are relevant but not "minimally necessary" for the interpretation of a relation. Supplementary material often appears as a parenthetical embedded within one of the arguments of a discourse relation and therefore we need to consider these cases as well to get the correct distribution of parentheticals over rhetorical relations. The problem with the annotation of supplementary material is that these units have not been annotated for discourse relations so we do not know what relation (if any) holds between a parenthetical of this type and its host (although we suspect in most cases these spans convey a relation similar to expansion). To give an example, (6) illustrates two cases of parentheticals annotated as supplementary material in Arg1 (6a) and Arg2 (6b) of a connective.

(6) a The trust, *which was created as part of Manville's bankruptcy-law reorganization to compensate victims of asbestos-related diseases,* ultimately expects to receive $2.5 billion from Manville, **but** its cash flow from investments has so far lagged behind its payments to victims. (0283)

b Last season, Hartford Stage director Mark Lamos mounted a production at Lincoln Center, **and** currently two other productions – *one just closed at the Old Globe in San Diego and another now at the Seattle Rep* – overlap with Mr. Boyd's. (0819)

There are 293 examples in the corpus where supplementary material appears embedded within one of the arguments of a connective. Supplementary material appears more frequently within arguments that contain a connective: there are 127 examples of Arg2-embedded supplementary material, compared to 76 Arg1-embedded cases. However, more work will have to be done to evaluate the parenthetical nature of supplementary material in the PDTB, and then to distinguish among the various syntactic types that are indeed parentheticals. Our intuition is that many of these parenthetical supplements will be quite compatible with the syntactic types noted in the RST.

We end this section by pointing out that the annotation of supplementary material also allows us to investigate the larger discourse context in which the parenthetical and its host appear (e.g. in (6a) the host of the parenthetical functions as Arg1 of the connective 'but'). Table 5 shows the distribution of discourse relations where one of the arguments contains supplementary material, broken down according to the semantic type of the connective. The table shows that there is no significant difference between discourse relations in terms of allowing parentheticals to occur in their arguments, however parentheticals do occur slightly more frequently in the clause that contains the connective (Arg2).

Relation	Sup within Arg1	Sup within Arg2
TEMPORAL	7	31
CONTINGENCY	15	28
COMPARISON	30	36
EXPANSION	24	32
TOTAL:	76	127

Table 5: Distribution of Supplementary material

5. Discussion

5.1. The Elaboration/Expansion Distribution

The relation types classed as *elaboration*, as used in the RST, or *expansion*, as used in the PDTB, are the most common types of relations found in the respective corpora. Since their definitions overlap considerably, we therefore consider them jointly under the rubric Elaboration/Expansion. In extracting parentheticals from the RST and the PDTB, we might ask therefore whether the set of relations that we obtained also show the same preponderance of elaboration- and expansion-type relations. We classify our data into two primary types - ELABORATION/EXPANSION and NON-ELABORATION/EXPANSION type relations. Since the PDTB treats attribution as a separate phenomenon, we leave out the ATTRIBUTION relation encountered in the RST to facilitate comparison.[3]

Table 6 shows the proportions of ELABORATION/EXPANSION and NON-ELABORATION/EXPANSION relations in the two corpora, culled from the numbers seen in Tables 1 and 4.

For the RST, we take subtypes of ELABORATION to belong to the ELABORATION/EXPANSION group, as well as EXAMPLE and RESTATEMENT since recapitulation and exemplification are clearly related to a notion of elaboration.

[3] ATTRIBUTION seems to be a different type of relation altogether, relating some agent to a rhetorical argument ("X said Y") rather than being a relation between two full-fledged discourse arguments.

	RST		PDTB	
ELAB/EXPANSION	elab-add	331		
	restatement	22	expansion	20
	example	19		
	elab-set-member	14		
	elab-gen-spec	11		
	TOTAL	**397 (73.4%)**	**TOTAL**	**20 (9.1%)**
NON-ELAB/EXPANSION	circumstance	58		
	purpose	35	temporal	103
	condition	20	contingency	53
	antithesis	18	comparison	43
	concession	13		
	TOTAL	**144 (26.6%)**	**TOTAL**	**199 (90.9%)**

Table 6: Elaboration/Expansion vs Non-Elaboration/Expansion relations

NON-ELABORATION/EXPANSION is the residual class consisting of a heterogeneous set of rhetorical relations (CIRCUMSTANCE, PURPOSE, CONDITION, ANTITHESIS, CONCESSION) but excluding ATTRIBUTION.

The data in the first column clearly supports the general observation that this type of relations predominate in discourse, with ELABORATION/EXPANSION relations outnumbering NON-ELABORATION/EXPANSION relations by a considerable ratio of 73% to 27%.

In the PDTB, EXPANSION belongs to the ELABORATION/EXPANSION group, and the remaining three types (TEMPORAL, CONTINGENCY and COMPARISON) fall into the NON-ELABORATION/EXPANSION type.

The numbers from the PDTB however show a striking paradox relative to the numbers seen in the first column of Table 6 for the RST. Here, ELABORATION/EXPANSION relations account for only 9.1% of parentheticals extracted, whereas the residual NON-ELABORATION/EXPANSION category makes up a much larger 90.9%.

How do we account for the discrepancy in the ELABORATION/EXPANSION numbers between the RST and the PDTB? The key to answering this question is to recognize that discourse relations are manifested syntactically in two different ways: i) a relation between an object and a proposition is realized through some NP-modification structure; and ii) relations between two full-fledged propositions are realized through S- or VP-modification. As we have noted for the RST (see Table 2), the distinction between ELABORATION/EXPANSION and NON-ELABORATION/EXPANSION parentheticals correspond quite neatly to these two syntactic types. ELABORATION/EXPANSION parentheticals are mostly postmodifiers of a nominal object in the host whereas NON-ELABORATION/EXPANSION parentheticals are mostly instantiated as some VP or S modification of their clausal hosts.

The PDTB takes a different approach to discourse annotation in that it recognizes only relations between two propositions, or more generally, two *abstract objects* (Asher, 1993), the kinds likely to be syntactically instantiated as VP- or S-modification.[4] Given our observations that ELABORATION/EXPANSION parentheticals occur mostly as NP-modifiers, it follows straightforwardly that we should find a much lower number of such relations in the PDTB, since these types would not have been annotated.

A follow-up question naturally arises: why is there such a paucity of ELABORATION/EXPANSION type parentheticals manifested as S- or VP-modifiers? As we noted previously, S- or VP-modifier structures involve a relation between two clausal elements. The kind of ELABORATION/EXPANSION relations which would involve two clauses are overwhelmingly of the *conjunctive* kind - which would be relations marked by a cue phrase such as *and* in the PDTB, or treated as JOINT or LIST type relations in the RST. But such conjunctive relations are quite incompatible with a parenthetical structure. By definition, parentheticals are used to express less salient material in the discourse and therefore are used when one argument (the parenthetical) is of less importance than the other (the host). NP-modifier structures such as relative clauses are particularly suitable for expressing such relations. On the other hand, paratactic structures such as conjunction usually require the conjoining of two equally prominent sister nodes, and it is unusual, if not quite impossible, to nest one clausal conjunct within another. Hence, we would not expect to find conjunctions expressed as parentheticals.[5]

In fact, of the parentheticals extracted from the RST, none are of the paratactic JOINT or LIST type. And the handful of ELABORATION/EXPANSION parentheticals that we do encounter in the PDTB are of a more complex and rare syntactic nature, a nuance that is only hinted at by the annotations. (7) shows one such parenthetical (shown in italics).

(7) a Some may have forgotten - *and some younger ones may never have experienced* - what it's like to invest during a recession. (wsj1623)

In this case, there is a more complex process of constituent-sharing than is indicated by the annotations. We have a clausal object ("what it's like to invest during a recession"), which is associated by the annotator only with the host ("Some may have forgotten") but not with the parenthetical ("and some younger ones may never have experienced"). But the clausal object is clearly linked to both the parenthetical and the host, so what we have here is a complex conjunction structure where two non-constituent NP(SUBJ)+verb are conjoined and this unusual conjunction shares a common NP(OBJ). This is quite possibly a borderline case of parenthetical although the annotation (and the dashes in the text) do suggest it.

The discourse annotation of NP-modifier type ELABORATION/EXPANSION relations in the RST is also not without

[4]The corpus does contain annotations of entity-based coherence through a category called Entity Relations (EntRel). Nevertheless, EntRels do not correspond to the Object-Proposition type relations found e.g., in relative clauses and annotated in the RST as discourse relations.

[5]This property of parentheticals has been discussed for example by (Scott and Souza, 1990)

issues. If a proposition is related only to some nominal object in a higher clause, a reasonable annotation strategy would be to simply link the parenthetical to the relevant nominal phrase. However, the segmentation of discourse units is not done at this level of granularity, so by convention, the host sentence is split up into two halves and nominal postmodifiers are linked to the part of the host *containing* the nominal phrase. Therefore, *structurally* at the discourse level, real cases of S- or VP-modifier relations, where two propositions are actually related, become virtually indistinguishable from NP-modifier relations, where an entity and a proposition are related. The only way to make this distinction is to sift through the various labels of rhetorical relations in the RST, and then falling back on the syntax to disambiguate them into NP-modifier or S-/VP-modifier types, as we have done.

In sum, from a *structural* point of view, the RST conflates two discourse-level phenomena - NP-modifier vs S- or VP-modifier types of relations - into one, whereas the PDTB recognizes the distinction but only annotates one of them!

5.2. Location and Distribution of Cue Phrases

Cue phrases are the most salient indicators of discourse relations in a piece of text. Therefore, to better understand the behavior of parentheticals in discourse, we also need to understand the characteristic behavior of cue phrases in this context. A corpus like the PDTB is most useful for such an endeavour. Here, cue phrases - known as discourse connectives - are taken to be discourse-level predicates which lexically anchor discourse relations. In other words, much of the annotation took place when there was an explicit discourse connective which signalled the presence of a discourse relation. The examples extracted for this study, in fact, were all discourse relations anchored by a lexicalized connective. But there is a striking imbalance in the distribution of the data. As seen in Table 4, out of 219 parentheticals extracted from the corpus, 207 were cases of Arg2-parentheticals, i.e., contained a discourse connective within the parenthetical.

Type of Connective	Arg1-Parenth.	Arg2-Parenth.	Total
Subordinating Conj.	0 (0%)	205 (99%)	205
Discourse Adverbial	12 (100%)	2 (1%)	14
TOTAL	12	207	219

Table 7: Nested constructions in PDTB, by connective type

Why should there be so many Arg2-parentheticals relative to Arg1-parentheticals? To answer this question, we note that there are two main grammatical types of discourse connectives which appear within parentheticals: i) subordinating conjunctions (e.g. while, because, since); and ii) adverbials (instead, however, meanwhile). Interestingly, the overwhelming majority of the pervasive Arg2-Parenthetical type are cases where Arg2 contains a subordinating conjunction (205 tokens out of 207, see Table 7). Subordinating conjunctions in fact never co-occur with Arg1 Parentheticals. In addition, discourse adverbials, even though they may appear in both Arg1- and Arg2-Parentheticals (12 and 2 times respectively), are comparatively rare and their infrequency means that they do not affect the skewered distribution seen in Table 7.

It appears then that subordinating conjunctions are the most "conducive" to parentheticals. This makes sense when we consider the grammatical nature of these connectives - a subordinating conjunction introduces an S-BAR subordinate clause (the parenthetical) which is *structurally integrated* into the matrix clause (the host). This structural integration provides for a number of structural transformations, including moving the subordinated clause medially or initially, which would otherwise not be possible. A parenthetical construction is "born" essentially when the subordinate clause is moved medially. Since Arg2 is the clause containing the connective (the subordinating conjunction), these would be Arg2-parentheticals. The converse does not hold however: to obtain Arg1 parentheticals when Arg2 contains the subordinating conjunction, the Arg2 subordinate clause would have to be split into two and then be transformed in such a way that it winds up nesting the Arg1 matrix clause, a completely unheard of structural transformation.

The case with discourse adverbials is more complex. Syntactically, an adverbial clause is not necessarily structurally integrated into a matrix clause in a subordinated structure but in most cases it introduces an independent new sentence. The relation of the discourse adverbial to a preceding Arg1 argument is therefore not syntactic, as was the case with subordinating conjunctions. In fact, (Webber et al., 2003) argue that the adverbial is *anaphoric* instead of structural, the notion being that the adverbial contains some integrated deictic particle which refers back to a preceding discourse argument. In other words, the discourse argument containing the adverbial (Arg2) needs to *linearly* precede the other argument (Arg1). This anaphoric property of discourse adverbials reduces the space of structural configurations that would be compatible with parenthetical constructions, as illustrated in Table 8. The first configuration (row i) shows the canonical non-parenthetical pattern, where Arg2 containing the adverbial simply follows Arg1. Among the parenthetical configurations, there is only one possible configuration of Arg2-Parentheticals (row ii). This configuration is rarely encountered, possibly because the adverbial here only refers to the first portion of its Arg1 antecedent (ARG1a) and not the entire Arg1 clause (ARG1a + ARG1b). Among the Arg1-Parentheticals, one of the configurations (Table 8, row iii) is impossible , since the adverbial here does not precede ARG1 at all. This leaves only the option in row iv) where the adverbial points to the nested ARG1 parenthetical.

	Configuration	Parenthetical Type	Observations
i)	ARG1 ARG2+adv	Not Parenthetical	Common
ii)	ARG1a ARG2+adv ARG1b	Arg2-Parenthetical	Rare
iii)	ARG2a+adv ARG1 ARG2b	Arg1-Parenthetical	Impossible
iv)	ARG2a ARG1 ARG2b+adv	Arg1-Parenthetical	Possible

Table 8: Configurations involving discourse adverbials. [6]

An analysis of the location of cue phrases is more difficult with the RST corpus, since the corpus does not distinguish between relations containing cue phrases and those that do not. However, we have reliable information for the parentheticals extracted from the RST to make another observa-

tion about the behavior of cue phrases: S- or VP-modifier type parentheticals tend to contain cue phrases, whereas NP-modifier parentheticals do not. From Table 2, we can see that a large majority of VP- or S-modifier parentheticals fall into the following syntactic types: cue+S, to-infinitive (where the "to" might be treated as a shortened form of the purpose cue-phrase "in order to"), cue+N, cue+Adj, cue+participial. NP-modifiers are mostly relative clauses, participial clauses, NPs, PPs or S's.

In sum, the combined use of the PDTB and the RST tells us quite a bit about the location of cue phrases: they generally occur with the parenthetical and not the host, and they appear with parentheticals which are VP- or S-modifiers. As a side rule, we could also surmise that generally these cue phrases are subordinating conjunctions. Discourse adverbial cue phrases might be more likely to appear in the host sentence, but the data is sparse and further study would be needed to understand the behavior of adverbials with parentheticals.

5.3. On the Generation of Parentheticals

Our corpus study provides information to design a generation strategy for several types of parenthetical constructions. The overwhelming majority of parentheticals annotated in the two corpora are syntactically related to their host sentences. We only found 9 examples of interpolations in the RST that consist of full sentences and in the PDTB, the 12 cases of Arg1-parentheticals containing discourse adverbs are the only ones that can perhaps be included in this category. A generator should therefore focus more on parentheticals that are syntactically related to their host.

There are two main groups of syntactically related parentheticals. The more popular group, ELABORATION/EXPANSION- type parentheticals should be generated most frequently, using syntactic constructions such as relative clauses, nominal postmodifiers with non-finite clauses and NPs. No cue phrases or discourse connectives should be generated in these cases. Conjunctive or paratactic types within the ELABORATION/EXPANSION group should also be avoided.

The other group, NON-ELABORATION/EXPANSION-type parentheticals should be generated less frequently and should always involve the use of a cue phrase. The syntactic types used for this group should be subordinate clauses, to-infinitives and perhaps PPs.

Assuming that the distribution of rhetorical relations in the input of the generator is similar to the distribution of rhetorical relations in our two corpora, the numbers in Table 3 should give an indication as to how often to realize each type of relation as a parenthetical.

There are several types of parentheticals that the present corpus study is missing because of the nature of the annotation guidelines of the corpora used. For example, in the RST corpus phrasal expressions beginning with prepositions or connectives that have ambiguous discourse cues are not segmented as EDUs (8a), and neither are reduced relative clauses that contain an adjective without a verbal element (8b):

(8) a But the technology, [while reliable,] is far slower than the widely used hard drives. (wsj1971)

b Each $5000 bond carries one warrant, [exercisable from Nov. 28, 1989, through Oct. 26, 1994] to buy shares at an expected premium of 2 1/2 % to the closing share price when terms are fixed Oct. 26. (wsj1161)

We expect that these constructions will behave similarly to subordinating conjunctions and relative clauses respectively. This hypothesis could be confirmed by incorporating these non-annotated constructions into the grammar of an NLG system according to the strategy described above, and evaluating the output of the system.

6. References

N. Asher. 1993. *Reference to Abstract Objects in English*. Kluwer, Dordrecht.

O. Bonami and D. Godard. 2007. Parentheticals in underspecified semantics: The case of evaluative adverbs. *Research on Language and Computation*, 5(4):391–413.

N. Burton-Roberts. 2005. Parentheticals. In E. K. Brown, editor, *Encyclopaedia of Language and Linguistics*. Elsevier Science, 2nd edition edition.

L. Carlson, D. Marcu, and M. E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10, Morristown, NJ, USA. ACL.

N. Dehe and Y. Kavalova, editors, 2007. *Parentheticals*, chapter Parentheticals: An introduction, pages 1–22. Linguistik aktuell 106. Amsterdam Philadelphia: John Benjamins.

A. Knott, 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, editors, *Text Representation: Linguistic and Psycholinguistic Aspects*, pages 181–196. Benjamins, Amsterdam.

W. C. Mann and S. A. Thompson. 1987. Rhetorical Structure Theory: A theory of text organization. Technical Report ISI/RS-87-190, Information Sciences Institute, June 1987.

PDTB-Group. 2008. The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania.

D. Scott and C. S. Souza. 1990. Getting the message across in RST-based text generation. In C. Mellish R. Dale M. Zock, editor, *Current Research in Natural Language Generation*, pages 31–56. Academic Press.

A. Siddharthan. 2002. Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Student Research Workshop, ACL*.

B. Webber, M. Stone, A. Joshi, and A. Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.