

A Uniform Syntax and Discourse Structure: the Copenhagen Dependency Treebanks

Daniel Hardt

DH.ITM@CBS.DK

*Department of IT Management
Copenhagen Business School
Copenhagen, Denmark*

Editors: Stefanie Dipper, Heike Zinsmeister, Bonnie Webber

Abstract

I present arguments in favor of the Uniformity Hypothesis: the hypothesis that discourse can extend syntax dependencies without conflicting with them. I consider arguments that Uniformity is violated in certain cases involving quotation, and I argue that the cases presented in the literature are in fact completely consistent with Uniformity. I report on an analysis of all examples in the Copenhagen Dependency Treebanks (CDT) involving violations of Uniformity. I argue that they are in fact all consistent with Uniformity, and conclude that the CDT should be revised to reflect this.

Keywords: Discourse, Syntax, Treebank

1. Introduction

The Copenhagen Dependency Treebanks (CDT) are unusual in that they contain annotation of both syntactic and discourse structure, using a single dependency graph (Buch-Kromann and Korzen, 2010). Underlying this approach is a strong hypothesis about the relation of syntax and discourse: namely, that they are subject to uniform well-formedness conditions. In other words, discourse structure should extend the syntax graph without conflicting with it. Other discourse annotation projects have not followed this approach: for example, The Penn Discourse Treebank has been annotated independently of the syntactic annotation of the same texts.

In this paper, I present arguments in favor of the hypothesis that discourse can extend syntax dependencies without conflicting with them. I call this the Uniformity Hypothesis. While the design of the CDT was motivated by the Uniformity Hypothesis, the actual annotation practice has been more flexible, allowing a mechanism for annotating discourse relations in a way that allows for violations of Uniformity. This flexibility was introduced in response to arguments in the literature against Uniformity; in particular that of (Dinesh et al., 2005). Here it is argued that Contrast relations sometimes conflict with syntactic relations when quotation is involved. I argue that the examples of interest here are in fact consistent with Uniformity, once the semantics of the Contrast relation is considered in more detail.

This raises the question of whether the flexibility in CDT annotation is ever needed, or whether its annotation could be made completely consistent with Uniformity. To answer this question, I report on a systematic analysis of the examples in CDT that have been annotated in violation of Uniformity. There are three discourse relations having at least 10 occurrences of such violations. I

argue that, in each case, the annotation can be replaced with one that does not violate Uniformity. Based on this, I argue that the CDT should be revised to reflect this.

In what follows, I begin with some background on the CDT, focusing on the annotation of syntax and discourse. Next, I consider the argument of (Dinesh et al., 2005) against Uniformity, arguing that the relevant examples are in fact consistent with Uniformity. I then turn to an empirical analysis of the examples in CDT that have been annotated to indicate a violation of Uniformity. I argue that all these cases are in fact best analyzed in a way that is consistent with Uniformity, and conclude that the CDT annotations should be modified to conform with Uniformity in all cases.

1.1 Background

The Copenhagen Dependency Treebanks, CDT, consist of five parallel open-source treebanks for Danish, English, German, Italian, and Spanish. The treebanks are being annotated manually with respect to syntax, discourse, anaphora, morphology, as well as translational equivalence (word alignment) between the Danish source text and the target texts in the four other languages. At this point it is primarily the Danish treebank that has been annotated for both discourse and syntax, so in this paper I will restrict attention to this treebank.

1.2 Syntax

The syntactic annotation of the CDT treebanks is based on the linguistic principles outlined in the dependency theory Discontinuous Grammar (Buch-Kromann, 2006) and the syntactic annotation principles described in (Kromann, 2003), (Buch-Kromann et al., 2007), and (Buch-Kromann et al., 2009). All linguistic relations are represented as directed labelled relations between words or morphemes. The model operates with a primary dependency tree structure in which each word or morpheme is assumed to act as a complement or adjunct to another word or morpheme, called the governor (or head), except for the top node of the clause or unit, typically the finite verb.

1.3 Discourse

Just as sentence structure can be captured by dependencies that link up the words and morphemes within a sentence, discourse structure can also be captured by dependencies that link up the words within an entire discourse. The CDT discourse annotation consists in linking up each clause’s top node with its nucleus (understood as the unique word within another clause that is deemed to govern the relation) and labelling the relations between the two nodes. The inventory of discourse relations in CDT is described in the CDT manual. It borrows heavily from other discourse frameworks, in particular Rhetorical Structure Theory, RST (Mann and Thompson, 1987; Taboada and Mann, 2006; Carlson, Lynn and Marcu, Daniel and Okurowski, Mary Ellen, 2001) and the Penn Discourse Treebank, PDTB (Webber, 2004; Dinesh et al., 2005; Prasad et al., 2007, 2008), as well as (Korzen, 2006, 2007), although the inventory had to be extended to accommodate the great variety of text types in the CDT corpus. The inventory allows relation names to be formed as disjunctions or conjunctions of simple relation names, to specify multiple relations or ambiguous alternatives. One of the most important differences between the CDT framework and other discourse frameworks lies in the way texts are segmented. In particular, CDT uses words as the basic building blocks in the discourse structure, while most other discourse frameworks use clauses as their atomic discourse

units, including RST, PDTB, GraphBank (Wolf and Gibson, 2005), and the Potsdam Commentary Corpus, PCC (Stede, 2004).

2. Uniformity of Syntax and Discourse

According to the hypothesis of Uniformity, the same well-formedness conditions apply to the discourse structure as to syntactic structure. In a very real sense CDT does not distinguish between syntax and discourse, since both discourse and syntax links are part of the same graph. I will focus on one particular aspect of Uniformity, which is easier to appreciate in terms of relations on bracketed structures. Consider the following structures S1 and S2

S1= [... [A] ...]

S2= [... [B] ...]

Given these structures, we allow a relation between S1 and S2, but we do not allow crossing relations involving embedded elements A or B: such as a link from A to B, from S1 to B, or from A to S2. Consider the following constructed example

- (1) John said it was raining. But it was not raining.

Here, we have

S1= [John said [it was raining]]

S2= [But it was not raining]]

where A (embedded within S1) is *it was raining*. One might be tempted to define a Contrast relation between A and S2. But Uniformity requires that the relation be between S1 and S2.

3. Quotation: an Apparent Violation of Uniformity

Dinesh et al. (2005) argue that just such violations of Uniformity can be observed in cases involving quotation. Consider Example 2:

- (2) The current distribution arrangement ends in March 1990, although Delmed said it will continue to provide some supplies of the peritoneal dialysis products to National Medical, the spokeswoman said. [(12) in (Dinesh et al., 2005)]

S1= [The current distribution arrangement ends in March 1990]

S2= [Delmed says [it will continue to provide some supplies of the peritoneal dialysis products to National Medical]...]

Here there is an embedded element in S2, which I call B: *it will continue to provide some supplies of the peritoneal dialysis products to National Medical*. (I ignore the final attribution to the spokeswoman.) Dinesh et al. argue that the discourse relation of Contrast, signalled by “although”, does not hold between S1 and S2, but between S1 and the embedded element B. According to Dinesh et al.: “*although* as a discourse connective denies the expectation that the supply of dialysis products will be discontinued when the distribution arrangement ends. It does not convey the expectation that Delmed will not say such things”. To evaluate this argument, we must look at the semantics of Contrast and quotation.

4. Semantics of Contrast

Dinesh et al. (2005) are assuming that Contrast typically involves a “denial of expectation”. This is a standard view of Contrast, and the Penn Discourse Treebank Annotation Manual (Prasad et al., 2007) defines Concession as a subtype of Contrast, characterized by “denial of expectation”, stating this:

The type Concession applies when the connective indicates that one of the arguments describes a situation A which causes C, while the other asserts (or implies) not C.

The RST Annotation Manual (Carlson, Lynn and Marcu, Daniel, 2001, p. 50) says this about Concession as a type of Contrast:

...a Concession relation is always characterized by a violated expectation (p 50)

But what exactly is meant by a “violated expectation”? In an influential early discussion, (Hobbs, 1985, p. 22), defines “Violated Expectation” as follows:

Infer P from the assertion of S_0 and $\neg P$ from the assertion of S_1 .

Hobbs illustrates this with the following example:

- (3) John is a lawyer, but he is honest.

From S_1 *John is a lawyer*, Hobbs argues, one can infer $P = \text{John is dishonest}$, while S_2 is *not P* (John is honest).

From this discussion, it is clear that Contrast between S_1 and S_2 normally involves a contradiction – S_1 implies some P and S_2 implies *not P*. At the same time, a normal, felicitous discourse must be logically consistent. What this means is that the contradiction in a Contrast must always be safely “packaged” to avoid an inconsistent discourse. A standard way of achieving this is that inferences from S_0 to P and from S_1 to *not P* are based on different background assumptions, which I will call *Back1* and *Back2*. A felicitous discourse does not require that one be committed to the truth of *Back1* and *Back2*, but, rather, one must be willing to temporarily entertain them.

We now return to Example 2. Recall that $S_1 = [\text{current distribution arrangement ends}]$, and $S_2 = [\text{Delmed says some supplies will continue}]$, with $B = [\text{some supplies will continue}]$. Dinesh et al. argue for Contrast between S_1 and B : I will call this **Case 1**. Uniformity would dictate Contrast between S_1 and S_2 , which I will call **Case 2**.

Following Hobbs’ analysis, we want to identify some P that we can infer from S_1 , such that *not P* can be inferred from B (in Case 1) or from S_2 (in Case 2). It is crucial that these two inferences rest on different background assumptions; otherwise the discourse would be inconsistent. Furthermore, a discourse is only felicitous if the background assumptions are salient and one is willing to entertain the possibility that they are true. I call the background assumptions underlying the inference to P and *not P* *Back1* and *Back2*, respectively.

In Case 1, the background assumption for inferring P , *Back1*, is this: *supplies only come from current distribution arrangement*. Together with S_1 , *current distribution ends*, one can infer P : *no*

supplies will continue. The background assumption for inferring *not P*, *Back2*, is empty, since *not P* is identical to B. Thus, I agree with Dinesh et al. that Contrast between S1 and B is coherent. However, Case 2 also supports Contrast. Here, we have S2 *Delmed says some supplies will continue* instead of B *some supplies will continue*. Instead of an empty background assumption *Back2*, we have *Delmed speaks truthfully*. From *Back2* and S2 one can infer *not P*. The two cases are shown below:

Case 1: Contrast S1,B

- Back1: supplies only come from current distribution arrangement
- S1: current distribution ends
- P: no supplies will continue
- Back2: <empty>
- B: some supplies will continue
- not P: some supplies will continue

Case 2: Contrast S1,S2

- Back1: supplies only come from current distribution arrangement
- S1: current distribution ends
- P: no supplies will continue
- Back2: Delmed speaks truthfully
- S2: Delmed says some supplies will continue
- not P: some supplies will continue

I have argued that Example 2 is in fact consistent with the uniformity hypothesis – A discourse relation of Contrast between the top-level constituents S1 and S2 is consistent with the notion that Contrast always requires conflicting material to be properly “packaged”, and quotations are one typical means for doing so.

Now, while the annotation approach in the CDT was originally motivated in part by the Uniformity Hypothesis, the actual annotation policy has been more flexible, incorporating a special Star notation for the express purpose of indicating the kinds of violations of Uniformity being discussed here. I have attempted to show that this argument is not convincing. There could of course be other cases in which Uniformity must be violated. Indeed, there are 135 examples in the CDT in which the Star notation has been used to indicate violations of uniformity. I now turn to an examination of these examples.

Relation	Total Occurrences	Occurrences with *	Percentage
CONJUNCTION	1938	76	3.921
CONTR	195	15	7.692
AGENTIVE	156	11	7.692
CONC	107	7	6.542
CONST	109	7	6.422
FORMAL	91	6	6.593
TELIC	150	6	4.000
TIME	64	3	4.687
EXPR	17	2	11.76
QUEST	48	2	4.166

Table 1: Occurrences of Star Notation in CDT

5. An Empirical Analysis of Uniformity in CDT

5.1 Use of the Star Notation in CDT

In CDT a relation can be written with a * either to the left or right (or both), to indicate that the left or right argument is embedded. Consider relation R linking S0 and S1, where

S0= [... [A] ...] and

S1= [... [B] ...]

If we have *R[S0,S1], this is a way of indicating R[A,S1], while R*[S0,S1] indicates R[S0,B]. In all cases I have observed, the embedded element referred to with the Star notation is *quoted material*. By quoted material, I mean to include both indirect and direct quotes, including sentential complements not only of verbs of saying but also verbs of belief. Thus I include *John said “it is raining”*, *John said that it is raining* and *John thinks that it is raining*. In all three cases I call *it is raining* the quoted material.

Table 1 gives the relations with which the Star notation occurs. Below I consider in detail the relations Contrast, Agentive, and Conjunction. The remaining relations have very small numbers of occurrences. I begin with a detailed consideration of Contrast, since it relates directly to the argument discussed above, which motivated the Star notation.

5.2 Contrast

I begin with Contrast. For each example, there are two top-level clauses, S1 and S2, and in each case the “*” notation has been used to indicate a Contrast relation where one of the arguments is embedded. Either S1 or S2 (or both) contains an embedded element, which is the complement of a saying verb. This gives rise to three possibilities. Table 2 gives the distribution of the Contrast examples with respect to these three possible categories.

For Category 1, recall Example 2. There, the argument was that Contrast was valid between S1 and S2, based on a background assumption that *Delmed speaks truthfully*. This argument can be made for all three categories. Assume that A = S1 or is embedded within S1, and B = S2 or is embedded within S2. In each case, the embedded element is the complement of “X SAYS”. If there

Category	S1	S2	Occurrences (File Id's)	Count
1	X SAYS A	B	688, 1182, 840, 465, 729, 138	6
2	A	X SAYS B	0005, 0366, 0654, 0986, 0683, 1251, 1014, 1214, 1527	9
3	X SAYS A	X SAYS B		0

Table 2: Occurrences of Star Notation with Contrast

is a Contrast relation between A and B, then, under the assumption that X is truthful, there must also be a Contrast relation between S1 and S2.¹

Below I consider an example of each category.

There are six examples in CDT in Category 1 “X SAYS A; B”. Consider this example (file 0688):

S1: [Administrerende direktør Peter Christoffersen siger, at [der hverken er forhandlinger eller sonderinger mellem Baltica og Skandia i øjeblikket].]

[CEO Peter Christoffersen says that [there are neither negotiations or explorations between Baltica and Skandia at the moment.].]

S2: [Skandia har tidligere ønsket et giftermål med Baltica.]

[Skandia had earlier wanted an alliance with Baltica.]

Simplifying a bit, we have A = *there are no negotiations* and S2 = *Skandia had wanted an alliance*.

The reasoning is directly parallel to that of Example 2: there is a “Violated Expectation” between A and S2 – in this case S2 (that an alliance was desired) sets up an expectation that there would be negotiations, and A violates that expectation. This is the Contrast relation annotated using the Star notation, indicating a relation between the embedded A with the top-level S2. However, under the assumption that the speaker, Peter Christoffersen, is truthful, then S1 can be inferred to violate the expectation just as A does.

Recall that Category 2 is “A; X SAYS B”. We have this example (file 0986):

S1= [De europæiske stålfabrikker befinder sig midt i den værste nedgang i ti år, uden udsigt til forbedring i år.]

[The European steel factories find themselves in the midsts of the worst downturn in ten years, with no prospects of improvement this year.]

S2= [[Men det er tid til at købe aktier i stålintustrien], siger erhvervsanalytikere.]

[[But this is the time to buy stock in the steel industry], say business analysts.]

The annotator found a contrast between S1 and B; since S1 describes a downturn in the steel industry, one could infer that this is not the time to buy stock in steel, while B expresses the opposite.

1. As an anonymous reviewer points out, Contrast does not always involve a denial of expectation; it can also involve a “juxtaposition of viewpoints” where two or more arguments of a given relation differ. It is worth mentioning that the Penn Discourse Treebank annotates two subtypes of Contrast, called *juxtaposition* and *opposition*. The logic of my argument concerning Contrast has been limited to cases involving denial of expectation. It is possible that this argument would not apply to examples involving these other types of Contrast, which would then support the argument that Uniformity cannot be maintained. No such cases of Contrast were found in the CDT however.

Category	S1	S2	Occurrences (File Id's)	Count
1	X SAYS A	B	0366, 1173, 0538, 1035	4
2	A	X SAYS B	0581, 0705, 0465, 1173, 0065, 1259	6
3	X SAYS A	X SAYS B	0001	1

Table 3: Occurrences of Star Notation with Agentive

Under the assumption that what business analysts say is true, there is also a contrast between S1 and S2.

5.3 Agentive (Cause/Reason)

The discourse relation Agentive in CDT is meant to indicate that one clause expresses a cause or reason for another clause. Table 3 gives the occurrences of the Star notation in connective with Agentive.

I begin with an example from Category 1 “X SAYS A; B” (file 1173):

S1 = [“Jeg respekterer virkelig Orlando,”] siger Michela Buscemi.]

[“I really respect Orlando,”] says Michela Buscemi.]

S1a= [To af hendes brødre er blevet dræbt af mafiaen, og hun vidnede i retten mod de mistænkte drabsmænd.]

[Two of her brothers were killed by the Mafia, and she testified in court against the suspected killers.]

S2= [“Han var den eneste, der rakte en hånd frem for at hjælpe mig.”]

[“He was the only one who came forward to help me.”]

Here we have A = “*I really respect Orlando*”, and B = “*He was the only one who came forward to help me*”.

(Note that the Agentive relation skips over the intervening sentence which I call here S1a.) Clearly in the reported discourse, the speaker Michela Buscemi is offering B as reason for A. However, in the discourse of the text, I argue that the writer is offering S2 as a reason for S1: that is, the fact that the speaker says she respects Orlando is explained by the fact that the speaker has a belief about him, namely that he helped her. It is perfectly coherent to see B (Orlando helped Buscemi) as a reason for A (Buscemi respects Orlando), but it is equally coherent to see S2 (Buscemi says Orlando helped her) as a reason for A. This latter view, I argue, is the correct view of the text, and this is what Uniformity would suggest.

More generally, the reasoning is parallel to that with Contrast: given that X is truthful, then if there is an Agentive relation between S1 and B there is also an Agentive relation between S1 and S2.

Category 2 “A; X SAYS B” (file 1259):

S1= [Og i hvert fald kan vi ikke måle kvaliteten af vore dages barndom ud fra de normer, der gjaldt dengang, vi selv var børn.]

[And in any case, we can’t measure the quality of today’s childhood based on the norms that were relevant when we were children.]

Category	S1	S2	Count
1	X SAYS A	B	32
2	A	X SAYS B	30
3	X SAYS A	X SAYS B	12

Table 4: Occurrences of Star Notation with Conjunction

S2= [”Børns vilkår har ændret sig så meget de seneste år, at vi faktisk ikke har noget sammenligningsgrundlag,”] siger han.]

[”The conditions of children have changed so much in recent years, that we actually don’t have a basis for comparison,”] he said.]

Here we have S1 = *And in any case, we can’t measure the quality . . .*, and B = *“The conditions of children have changed so much . . .”*.

It may well be that the quoted speaker is offering B as a *reason* for S1. This is presumably why the annotator used the Right-Star notation to indicate B as the second argument for the Agentive relation. But it is equally reasonable to argue that the writer is offering S2 as a *reason* for S1. Just as in previous examples: if the speaker is willing to entertain the assumption that X is truthful, then the fact that B is a reason for A means that X SAYS B is a reason for A.

Category 3

X SAYS A; X SAYS B

S1= [De hævder, at [Ruslands vej til demokrati går gennem diktatur.]]

[They claim that [Russia’s path to democracy goes through dictatorship.]]

S2= [I en af deres artikler hedder det: [”I et autoritært regime lagdel samfundet og forskellige interesser modnes.]]

[In one of their articles, it is stated: [”In an authoritarian regime, society is stratified and different interests matured.”]]

Here the annotator used both a left and right Star, indicating an Agentive relation between A and B. The statement in B, *In an authoritarian regime, . . .* provides an explanation for A, that Russia’s path to democracy goes through dictatorship. But this also supports an Agentive relation between S1 and S2: the author of the text is offering S2 as a reason for S1. That is, the reason the speakers (*they*) make the claim A, is that they have the beliefs B.

5.4 Conjunction

Table 4 shows the distribution of Star-notated Conjunction relations with respect to the different categories.

Because of the large number of Conjunction occurrences with the Star notation, I do not list the specific file id’s, as I did with Agentive and Contrast. Furthermore, Conjunction does not place specific semantic demands on its arguments in the way that Contrast and Agentive does. The general argument concerning Conjunction is the same: assume A is S1 or embedded within S1, and B is S2 or embedded within S2. If Conjunction holds between A and B, then assuming X is truthful, Conjunction must hold between S1 and S2.

The following is an example from Category 1, “X SAYS A, B” (file 1259):

S1= [[-Jeg har taget noget tøj med til Camilla,] forklarede Bjørn, da de var kommet ind i stuen.
]

[[I have brought some clothes for Camilla,] explained Bjørn, when they had come into the living room.]

S2= [-Du kan bare hente noget mere, hvis der ikke er nok.]

[-You can just get some more, if there is not enough.]

We have A = *I have brought some clothes for Camilla* and B = *-You can just get some more, if there is not enough.*. The dashes are meant to indicate direct quotations. So it is implicit that the speaker of S1, Bjørn, is also the speaker of S2. Thus S2 could be preceded with *And Bjørn continued:* without change to the meaning. The felicity of the connective *and* supports my claim that the Conjunction relation need not apply to the embedded A, but can relate the top level clauses S1 and S2.

6. Discussion

The CDT was originally formulated in accordance with Uniformity; subsequently, the Star notation was added because it was felt that it was necessary to violate Uniformity in certain cases. In this paper, I have argued that this is not the case. There are three relations with at least ten occurrences of the Star notation in CDT: Contrast, Agentive and Conjunction.² For Contrast and Agentive, I have made a general argument that quotations always involve a temporary assumption that the speaker is truthful and therefore the semantic relation attributed to the speaker will also be attributable to the writer. Because of this I argued that quotations involving Contrast or Agentive do not require a violation of Uniformity, based on an examination of all the relevant examples. The general argument involving Conjunction is similar, and furthermore Conjunction places rather weak requirements on its two arguments. Based on these arguments, I have argued that Uniformity can indeed be maintained in CDT.

I have not attempted to argue that these notations are *better* than those which violate Uniformity. For example, an anonymous reviewer points to the following pattern: John says X but John says Y, where it might well be that the fundamental contrast is between X and Y – that is, annotating a Contrast relation between X and Y might intuitively be the best choice. I do not deny that this might well be the case – I have merely attempted to show that it is possible to rule out such annotations and still be able to find acceptable annotations for all the data of the Danish portion of the CDT.

7. Conclusion

While most treebank work has focused on syntax, there is a growing interest in treebanks that involve annotation of discourse structure. The CDT is unusual in that discourse structure is treated as an extension of syntax, with an underlying assumption of Uniformity – that discourse and syntax can be annotated as part of single well-formed graph. Other major discourse treebank projects have not followed this approach, instead annotating discourse independently of syntax, reflecting

2. There is no reason to assume that apparent Uniformity violations are confined to these three relations. For example, an anonymous reviewer points out that the PDTB contains apparent Uniformity violations with a variety of other relations, including Temporal and Contingency. Whether these apparent violations can also be removed is a topic for future investigation.

a widespread view that Uniformity cannot be maintained, and the CDT was recently modified to allow violations of Uniformity.

In this paper I have shown that the data of the CDT can in fact be annotated consistently with Uniformity, and I have concluded that the CDT can and should be modified to remove all conflicts between discourse and syntax. In future work, I intend to investigate larger discourse treebanks such as the Penn Discourse Treebank to further investigate the possibility of maintaining Uniformity. This is less straightforward than the current investigation, since the deviations from Uniformity are not explicitly marked in the Penn Discourse Treebank, but it will provide a much richer empirical basis for exploring the hypothesis of Uniformity.

References

- Matthias Buch-Kromann. *Discontinuous Grammar. A dependency-based model of human parsing and language learning*. Copenhagen Business School, Copenhagen, 2006.
- Matthias Buch-Kromann and Iørn Korzen. The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks. In *Proceedings of ACL Linguistic Annotation Workshop*, 2010.
- Matthias Buch-Kromann, Jürgen Wedekind, and Jakob Elming. The Copenhagen Danish-English Dependency Treebank v. 2.0. <http://code.google.com/p/copenhagen-dependency-treebank>, 2007.
- Matthias Buch-Kromann, Iørn Korzen, and Henrik Høeg Müller. Uncovering the ‘lost’ structure of translations with parallel treebanks. *Copenhagen Studies in Language*, (38):199–224, 2009.
- Carlson, Lynn and Marcu, Daniel. Discourse tagging reference manual. Technical Report ISI-TR-545, ISI, September 2001.
- Carlson, Lynn and Marcu, Daniel and Okurowski, Mary Ellen. Building a discourse tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech*, Aalborg, Denmark, 2001.
- Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Attribution and the (non-)alignment of syntactic and discourse arguments of connectives. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 29–36, 2005.
- Jerry Hobbs. On the coherence and structure of discourse. Report CSLI-85-37, Center for the Study of Language and Information, Stanford University, 1985.
- Iørn Korzen. Endocentric and exocentric languages in translation. *Perspectives: Studies in Translationology*, 13(1):21–37, 2006.
- Iørn Korzen. Linguistic typology, text structure and appositions. In Iørn Korzen, Marie Lambert, and Hélène Vassiliadou, editors, *Langues d’Europe, l’Europe des langues. Croisements linguistiques*, volume 22, pages 21–42. Scolia, 2007.
- Matthias T. Kromann. The Danish Dependency Treebank and the DTAG treebank tool. In *Treebanks and Linguistic Theories (TLT 2003)*, page 217–220, Växjö, 2003.

- William C. Mann and Sandra A. Thompson. Rhetorical structure theory. A theory of text organization. ISI/RS-87-190, ISI: Information Sciences Institute, 1987.
- Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. The Penn Discourse TreeBank 2.0. Annotation Manual. The PDTB Research Group, 2007.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings 6th Int. Conf. on Language Resources and Evaluation*, Marrakech, Morocco, 2008.
- Manfred Stede. The Potsdam Commentary Corpus. In *Proceedings of ACL 2004 Workshop on Discourse Annotation*, pages 96–102, 2004.
- Maite Taboada and William C. Mann. Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459, 2006.
- Bonnie Webber. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science*, 28:751–779, 2004.
- Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287., 2005.