

# The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks

Matthias Buch-Kromann

Iørn Korzen

Center for Research and Innovation in Translation and Translation Technology  
Copenhagen Business School

## Abstract

We propose a unified model of syntax and discourse in which text structure is viewed as a tree structure augmented with anaphoric relations and other secondary relations. We describe how the model accounts for discourse connectives and the syntax-discourse-semantics interface. Our model is dependency-based, ie, words are the basic building blocks in our analyses. The analyses have been applied cross-linguistically in the Copenhagen Dependency Treebanks, a set of parallel treebanks for Danish, English, German, Italian, and Spanish which are currently being annotated with respect to discourse, anaphora, syntax, morphology, and translational equivalence.

## 1 Introduction

The Copenhagen Dependency Treebanks, CDT, consist of five parallel open-source treebanks for Danish, English, German, Italian, and Spanish.<sup>1</sup> The treebanks are 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.

The treebanks build on the syntactic annotation in the 100,000-word Danish Dependency Treebank (Kromann 2003) and Danish-English Parallel Dependency Treebank (Buch-Kromann *et al.* 2007). Compared to these treebanks, which are only annotated for syntax and word alignment, the new treebanks are also annotated for discourse, anaphora, and morphology, and the syntax annotation has been revised with a much more fine-grained set of adverbial relations and a number of other adjustments. The underlying Danish PAROLE text corpus (Keson and Norling-Christensen 1998) consists of a broad mixture of 200-250 word excerpts from general-purpose texts.<sup>2</sup> The texts were translated into the

other languages by professional translators who had the target language as their native language.

The final treebanks are planned to consist of approximately 480 fully annotated parallel texts for Danish and English, and a subset of approximately 300 fully annotated parallel texts for German, Italian, and Spanish, with a total of approximately 380,000 ( $2 \cdot 100,000 + 3 \cdot 60,000$ ) annotated word or punctuation tokens in the five treebanks in total. So far, the annotators have made complete draft annotations for 67% of the texts for syntax, 40% for word alignments, 11% for discourse and anaphora, and 3% for morphology. The annotation will be completed in 2010.

In this paper, we focus on how the CDT treebanks are annotated with respect to syntax and discourse, and largely ignore the annotation of anaphora, morphology, and word alignments. In sections 2 and 3, we present the syntax and discourse annotation in the CDT. In section 4, we present our account of discourse connectives. In section 5, we briefly discuss the syntax-discourse-semantics interface, and some criticisms against tree-based theories of discourse.

## 2 The syntax annotation of the CDT

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

<sup>1</sup>The treebanks, the annotation manual, and the relation hierarchy can be downloaded from the web site:

<http://code.google.com/p/copenhagen-dependency-treebank>

<sup>2</sup>In practice, the use of text excerpts has not been a problem for our discourse annotation: we mainly annotate text ex-

cerpts that have a coherent discourse structure, which includes 80% of the excerpts in our text corpus. Moreover, given the upper limit on the corpus size that we can afford to annotate, small text excerpts allow our corpus to have a diversity in text type and genre that may well offset the theoretical disadvantage of working with reduced texts.

of the sentence or unit, typically the finite verb. This structure is augmented with secondary relations, e.g., between non-finite verb forms and their subjects, and in antecedent-anaphor relations. Primary relations are drawn above the nodes and secondary below, all with directed arrows pointing from governor to dependent. The relation label is written at the arrow tip, or in the middle of the arrow if a word has more than one incoming arrow.

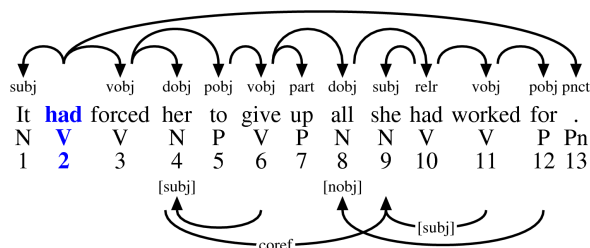


Figure 1. Primary dependency tree (top) and secondary relations (bottom) for the sentence “It had forced her to give up all she had worked for”.

An example is given in Figure 1 above. Here, the arrow from “had<sub>2</sub>” to “It<sub>1</sub>” identifies “It” as the subject of “had”, and the arrow from “forced<sub>3</sub>” to “to<sub>5</sub>” identifies the phrase headed by “to” as the prepositional object of “forced”. Every word defines a unique phrase consisting of the words that can be reached from the head word by following the downward arrows in the primary tree.<sup>3</sup> For example, in Figure 1, “worked<sub>11</sub>” heads the phrase “worked<sub>11</sub> for<sub>12</sub>”, which has a secondary noun object *nobj* in “all<sub>8</sub>”; “had<sub>10</sub>” heads the phrase “she<sub>9</sub> had<sub>10</sub> worked<sub>11</sub> for<sub>12</sub>”; and “It<sub>1</sub>” heads the phrase “It<sub>1</sub>”. Examples of secondary dependencies include the coreferential relation between “her<sub>4</sub>” and “she<sub>9</sub>”, and the anaphoric relation in Figure 2. Part-of-speech functions are written in capital letters under each word. The inventory of relations is described in detail in our annotation manual (posted on the CDT web site).

Dependency arrows are allowed to cross, so discontinuous word orders such as topicalisations and extrapositions do not require special treatment. This is exemplified by the discontinuous dependency tree in Figure 2, in which the relative clause headed by “was<sub>7</sub>” has been extraposed from the direct object and placed after the time adverbial “today<sub>5</sub>”.<sup>4</sup>

<sup>3</sup>Because of this isomorphism between phrases and head words, a dependency tree can always be represented as a phrase-structure tree in which every phrase has a unique lexical head; the resulting phrase-structure tree is allowed to contain crossing branches.

<sup>4</sup>In our current syntax annotation, we analyze the initial connective or conjunction as the head of the subordinate clause;

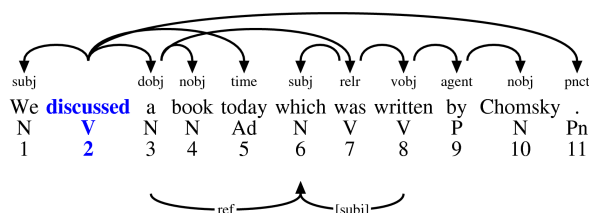


Figure 2. Primary dependency tree and secondary relations for the sentence “We discussed a book today which was written by Chomsky”.

Buch-Kromann (2006) provides a detailed theory of how the dependency structure can be used to construct a word-order structure which provides fine-grained control over the linear order of the sentence, and how the dependency structure provides an interface to compositional semantics by determining a unique functor-argument structure given a particular modifier scope (ie, a specification of the order in which the adjuncts are applied in the meaning construction).<sup>5</sup>

### 3 The discourse annotation of the CDT

Just like sentence structures can be seen as dependency structures that link up the words and morphemes within a sentence (or, more precisely, the phrases headed by these words), so discourse structures can be viewed as dependency structures that link up the words and morphemes within an entire discourse. In Figures 1 and 2, the top nodes of the analysed sentences (the only words without incoming arrows) are the finite verbs “had<sub>2</sub>” and “discussed<sub>2</sub>” respectively, and these are shown in boldface. Basically, the CDT discourse annotation consists in linking up each such sentence top node with its nucleus (understood as the unique word within another sentence 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 et al, 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

in relative clauses, the relative verb functions as the head, i.e., the arrow goes from “a (book)” to “was (written)”.

<sup>5</sup>In terms of their formal semantics, complements function as arguments to their governor, whereas adjuncts function as modifiers; i.e., semantically, the governor (type X) acts as an argument with the modifier (type X/X) as its functor.

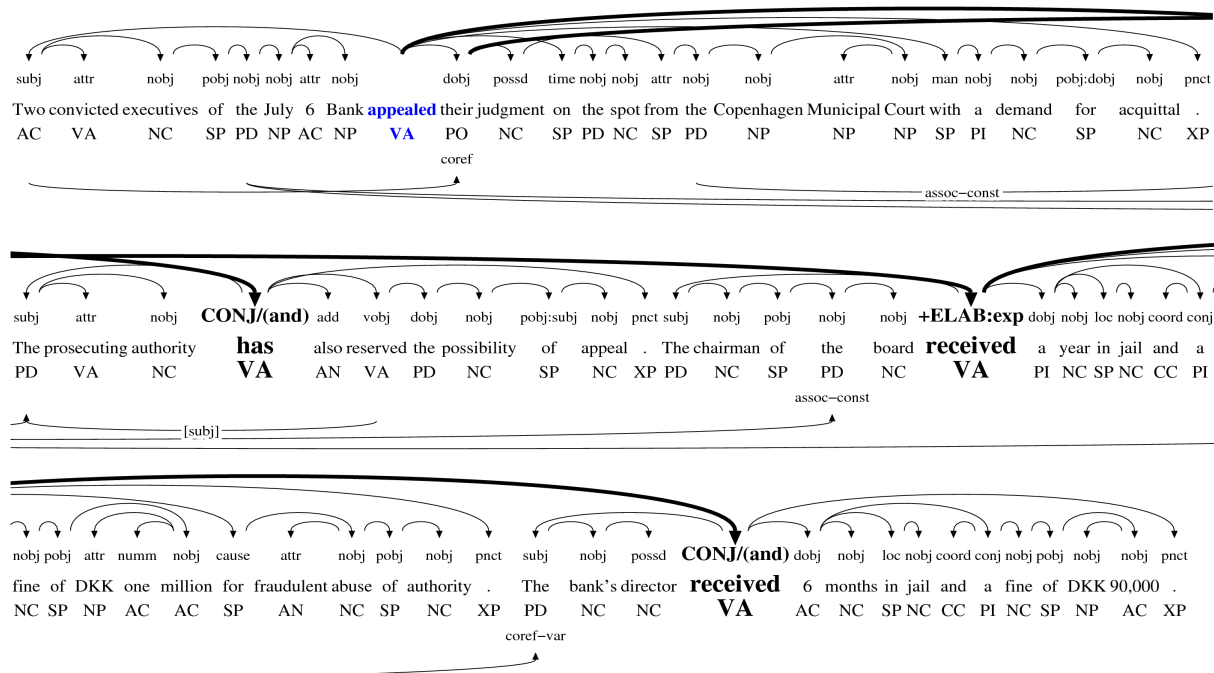


Figure 3. The full CDT analysis of (1) wrt. syntax, discourse, and anaphora.

variety of text types in the CDT corpus other than news stories. 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 2009).<sup>6</sup> This allows the nucleus and satellite in a discourse relation to be identified precisely by means of their head words, as in the example (1) below from the CDT corpus, where the second paragraph is analyzed as an elaboration of the deverbal noun phrase “their judgment” (words that are included in our condensed CDT analysis in Figure 4 are indicated with boldface and subscripted with numbers that identify them):

<sup>6</sup>As noted by Carlson and Marcu (2001), the boundary between syntax and discourse is rather unclear: the same meaning can be expressed in a continuum of ways that range from clear discourse constructions (“He laughed. That annoyed me.”) to clear syntactic constructions (“His laugh annoyed me.”). Moreover, long discourse units may function as objects of attribution verbs in direct or indirect speech, or as parenthetical remarks embedded within an otherwise normal sentence. CDT’s use of words as basic building blocks, along with a primary tree structure that spans syntax and discourse, largely eliminates these problems.

- (1) Two convicted executives of the July 6 Bank **appealed<sub>1</sub>** their<sub>2</sub> judgment on the spot from the Copenhagen Municipal Court with a demand for acquittal. The prosecuting authority **has<sub>3</sub>** also reserved the possibility of appeal.

The chairman of the board **received<sub>4</sub>** a year in jail and a fine of DKK one million for fraudulent abuse of authority [...]. The bank’s director **received<sub>5</sub>** 6 months in jail and a fine of DKK 90,000. (Text 0531)

The full CDT analysis of (1) is given in Figure 3, a more readable condensed version in Figure 4. The last sentence of the first paragraph, “The prosecuting authority has<sub>3</sub> also reserved the possibility of appeal”, is a conjunct to the first sentence, and its top node “has<sub>3</sub>” is linked to the top node of the first sentence, “appealed<sub>1</sub>”. The slash after a relation name indicates an explicit or implicit discourse connective used by the annotators to support their choice of relation type.

As in CDT’s syntax annotation, the primary syntax and discourse relations must form a tree that spans all the words in the text, possibly supplemented by secondary relations that encode anaphoric relations and other secondary dependencies. Apart from this, CDT does not place any restrictions on the relations; in particular, a word

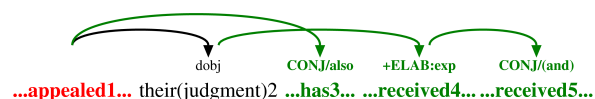


Figure 4. Condensed version of Figure 3.

may function as nucleus for several different satellites, discourse relations may join non-adjacent clauses, and are allowed to cross; and secondary discourse relations are used to account for the distinction between story line level and speech level in attributions.

#### 4 Discourse connectives

Discourse connectives play a prominent role in PDTB, and inspire the analysis of connectives in CDT. However, there are important differences in analysis, which affect the way discourse structures are construed. In a construction of the form “ $X C Y$ ” where  $X$  and  $Y$  are clauses and  $C$  is a discourse connective (such as “because”, “since”, “when”), three dependency analyses suggest themselves, as summarized in Table 5.




	Head	Conjunction	Marker
Syntax			
Semantics	$C(X, Y)$	$[C(Y)](X)$	$[Y(C)](X)$

Table 5. Three analyses of discourse connectives.

When analyzed as the head of the construction,  $C$  takes  $X$  and  $Y$  as its (discourse) complements; semantically, the meaning  $C'$  of  $C$  acts as functor, and the meanings  $X', Y'$  of  $X, Y$  act as arguments of  $C'$ . When analyzed as a subordinating conjunction,  $C$  subcategorizes for  $Y$  and modifies  $X$ ; semantically,  $C'$  computes a meaning  $C'(Y)$  from  $Y'$ , which acts as functor with  $X'$  as argument. Finally, analyzed as a marker,  $C$  modifies  $Y$  which in turn modifies  $X$ ; semantically,  $Y'$  selects its meaning  $Y'(C')$  based on the marker  $C'$  (i.e., the marker merely helps disambiguate  $Y$ );  $Y'(C')$  then acts as functor with argument  $X'$ .

The three analyses are markedly different in terms of their headedness, but quite similar in terms of their semantics. CDT opts for the marker analysis, with the obvious benefit that there is no need to postulate the presence of a phonetically empty head for implicit connectives. This analysis also implies that since discourse markers always modify the satellite, explicit and implicit discourse markers can be used to determine the discourse relation and its direction.

It is interesting that almost all theories of discourse structure, including RST, PDTB, GraphBank, PCC, and the dependency-based discourse analysis proposed by Mladová (2008), seem to analyze connectives as heads – even in the case where  $C+Y$  is an adverbial clause modifying  $X$ ,

where virtually all mainstream theories of syntax opt for one of the two other analyses. Perhaps current theories of discourse structure perceive discourse structure as a semantic rather than syntactic structure. In any case, it is not clear that this is the most fruitful analysis. A clear distinction between syntactic structure and semantic structure has proved crucial to the understanding of headedness in syntax (e.g. Croft 1995, Manning 1995), and it is one of the hardwon insights of syntax that semantic centrality or prominence is not directly reflected in the syntactic surface structure. Something similar might be true for discourse structure as well.

#### 5 Syntax-discourse-semantics interface

CDT models discourse structure as a primary dependency tree supplemented by secondary relations. We believe that a tree-based view of discourse provides many important benefits, most importantly a clear interface to syntax and compositional semantics. There has been several attempts to refute the tree hypothesis on empirical grounds, though, including Wolf and Gibson (2005), Prasad et al (2005), Lee et al (2008), and Stede (2009), who have put forward important criticisms. Our framework addresses many of these objections, including the many problems related to attribution verbs, which do require a complicated treatment in our framework with secondary dependencies. A full discussion of this topic is, however, beyond the scope of this paper.

#### 6 Conclusion

In this paper, we have presented a dependency-based view of discourse and syntax annotation where the syntax and discourse relations in a text form a primary dependency tree structure linking all the words in the text, supplemented by anaphoric relations and other secondary dependencies. The framework forms the basis for the annotation of syntax, discourse, and anaphora in the Copenhagen Dependency Treebanks. In future papers, we will address some of the criticisms that have been raised against tree-based theories of discourse.

#### 7 Acknowledgments

This work was supported by a grant from the Danish Research Council for the Humanities. Thanks to Bonnie Webber, Henrik Høeg Müller, Per Anker Jensen, Peter Colliander, and our three reviewers for their valuable comments.

## References

- Matthias Buch-Kromann 2006. *Discontinuous Grammar. A dependency-based model of human parsing and language learning*. Copenhagen: Copenhagen Business School.
- Matthias Buch-Kromann, Iørn Korzen, and Henrik Høeg Müller. 2009. Uncovering the 'lost' structure of translations with parallel treebanks. *Copenhagen Studies in Language* 38: 199-224.
- Matthias Buch-Kromann, Jürgen Wedekind, and Jakob Elming. 2007. The Copenhagen Danish-English Dependency Treebank v. 2.0. <http://code.google.com/p/copenhagen-dependency-treebank>
- Lynn Carlson and Daniel Marcu. 2001. *Discourse Tagging Reference Manual*. ISI Technical Report ISI-TR-545.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proc. of the 2nd SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics: 1-10.
- William Croft. 1995. What's a head? In J. Rooryck and L. Zaring (eds.), *Phrase Structure and the Lexicon*. Kluwer.
- Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives. *Proc. of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pp. 29-36.
- Britt Keson and Ole Norling-Christensen. 1998. PA-ROLE-DK. The Danish Society for Language and Literature.
- Iørn Korzen. 2006. Endocentric and Exocentric Languages in Translation. *Perspectives: Studies in Translatology*, 13 (1): 21-37.
- Iørn Korzen. 2007. Linguistic typology, text structure and appositions. In I. Korzen, M. Lambert, H. Vasiliadou. *Langues d'Europe, l'Europe des langues. Croisements linguistiques*. *Scolia* 22: 21-42.
- Matthias T. Kromann. 2003. The Danish Dependency Treebank and the DTAG treebank tool. In *Proc. of Treebanks and Linguistic Theories (TLT 2003)*, 14-15 November, Växjö. 217-220.
- Alan Lee, Rashmi Prasad, Aravind Joshi, and Bonnie Webber 2008. Departures from Tree Structures in Discourse: Shared Arguments in the Penn Discourse Treebank. *Proceedings of the Constraints in Discourse III Workshop*.
- William C. Mann and Sandra A. Thompson 1987. *Rhetorical Structure Theory. A Theory of Text Organization*. ISI: Information Sciences Institute, Los Angeles, CA, ISI/RS-87-190, 1-81.
- Christopher D. Manning. 1995. Dissociating functor-argument structure from surface phrase structure: the relationship of HPSG Order Domains to LFG. Ms., Carnegie Mellon University.
- Lucie Mladová, Šarka Zikánová, and Eva Hajičová. 2008. From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. In *Proc. 6th International Conference on Language Resources and Evaluation (LREC 2008)*.
- Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. 2007. The Penn Discourse TreeBank 2.0. Annotation Manual. The PDTB Research Group. <http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf>
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proc. 6th Int. Conf. on Language Resources and Evaluation*, Marrakech, Morocco.
- Manfred Stede. 2008. Disambiguating Rhetorical Structure. *Research on Language and Computation* (6), pp. 311-332..
- Maite Taboada and William C. Mann. 2006a. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies* 8/3/423.
- Maite Taboada and William C. Mann. 2006b. Applications of Rhetorical Structure Theory. *Discourse Studies* 8/4/567. <http://dis.sagepub.com>
- Bonnie Webber. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science* 28: 751-779.
- Bonnie Webber. 2006. Accounting for Discourse Relation: Constituency and Dependency. M. Dalrymple (ed.). *Festschrift for Ron Kaplan*. CSLI Publications.
- Florian Wolf and Edward Gibson 2005. Representing Discourse Coherence: A Corpus-Based Study. *Computational Linguistics* 31(2), 249-287.