Syntax-Centered and Semantics-Centered Views of Discourse. Can They be Reconciled?

Matthias Buch-Kromann, Daniel Hardt, and Iørn Korzen
Copenhagen Business School

Abstract

In this paper, we argue that there are two seemingly incompatible perceptions of discourse structure: a semantics-centered view and a syntax-centered view. In the semantics-based view, discourse structure is viewed as a structure that identifies the most important portions of the text and describes how they combine semantically. In the syntax-based view, discourse structure is viewed as an extension of syntax to the discourse level, which essentially links the syntactic trees for the individual sentences into one big tree structure. We will argue that these differences in perception may explain some of the central disagreements in the literature about the nature of discourse structure, in particular whether discourse structure is best viewed as a tree or a general graph. However, the two views are not as incompatible as they may seem at first sight, since the semantic discourse structure can be reinterpreted as a functor-argument structure that is derived from the syntactic tree structure. We describe the ramifications of the two views for the analysis of discourse markers, which are the focus of the discourse annotation in the Penn Discourse Treebank, and show how the syntax-based view can maintain a tree structure even for examples that seem to exhibit non-tree like properties in a semantics-based view.

1 Introduction

Most research on discourse structure builds on the premise that coherent texts have an associated internal structure that places constraints on how the meaning of the whole text is computed from the meanings of its individual clauses and sentences, and how the individual clauses and sentences are presented in the linear order. When texts appear coherent and easily comprehensible to readers, it is because they have a well-formed discourse structure that respects the listeners' conventions about linear order and sensible semantic and pragmatic interpretation, whereas texts that lack this property are perceived as being incoherent and difficult to comprehend. Discourse structure is mostly viewed as a tree structure, or at least a very tree-like structure, where the most important portions of the text are assumed to be located at or near the top of the tree, and the deepest parts of the tree are supposed to encode supplementary information that is less central to the writer's purposes and can be more easily excluded from a summary of the text. The individual branches in the discourse tree define discourse units which are supposed to form coherent textual units that can be interpreted in isolation, a property that can be used in reverse to identify the discourse units in a text. To varying degrees, this general framework forms the theoretical basis for discourse theories like Rhetorical Structure Theory (Mann and Thompson, 1987), the Linguistic Discourse Model (Polanyi, 1988), and Segmented Discourse Representation Theory (Lascarides and Asher, 2007), and for discourse treebanks like the English RST treebank (Carlson et al., 2001), the Discourse Graphbank (Wolf and Gibson, 2005), the Penn Discourse Treebank (Prasad et al., 2008) and related discourse treebanks (Mladová et al., 2008; Aktaş et

al., 2010), the Potsdam Commentary Corpus (Stede, 2008), and the Copenhagen Dependency Treebanks (Buch-Kromann and Korzen, 2010).

This mainstream view of discourse structure is to a very large degree inspired by the success with which syntactic theory has managed to account for intra-sentential structure. In mainstream syntax, the structure of a sentence is modelled by means of a tree augmented with additional structure which may be used to handle semantics or deal with non-canonical word order and secondary dependencies[1] (e.g., in topicalizations, control constructions, and relative clauses); this is true for a wide range of syntactic theories, including Head-Driven Phrase Structure Grammar[2] (Pollard and Sag, 1994), Lexical-Functional Grammar (Dalrymple et al., 1994), Government and Binding Theory (Chomsky, 1981/1993), Combinatory Categorial Grammar (Steedman, 2000), Tree-Adjoining Grammar (Joshi and Schabes, 1997), and different versions of dependency grammar (Hudson, 2010; Mel'čuk, 1988; Sgall et al., 1986; Duchier, 2001; Buch-Kromann, 2009; and many others). It is therefore tempting to try to reuse these mechanisms for the analysis of discourse, which is what most theories of discourse have sought to do (with Wolf and Gibson (2005) as the clearest exception). In syntax, there seems to be agreement about the general mechanisms needed to account for syntactic structure, although the specific implementational details vary greatly between the frameworks; but in discourse, there is a much lower level of agreement about the detailed theoretical interpretation and function of discourse structure and its relationship to syntax and discourse semantics.

In this paper, we seek to clarify some of these interpretational problems. By drawing on the insights and mechanisms from syntax and its relationship to sentential semantics, we hope to shed light on ways in which these insights may be carried over to our understanding of discourse. The paper is structured as follows. In section 2, we describe the blurry syntax-discourse boundary and the implications for the relationship between syntactic structure and discourse structure. In section 3, we describe the syntactic distinction between constituent structure and functor-argument structure, and argue in section 4 that this distinction is relevant for discourse as well. In section 5, we discuss the implications for the analysis of discourse connectives. In section 6, we argue that attribution is a particularly hard problem for a tree-based analysis of discourse, but that the problem can be resolved by either a more careful semantic analysis or a small extension of the compositional semantics. In section 7, we revisit some of the counter-examples that have been used to argue against a tree-based view of discourse structure. In Section 8, we identify some of the outstanding problems in a syntax-based view of discourse. In section 9, we describe how these insights have informed the syntax-based discourse annotation in the Copenhagen Dependency Treebanks. Our conclusions are presented in section 10.

---

1 By a secondary dependency we mean the phenomenon that a single phrase may sometimes function as a complement or adjunct in several phrases simultaneously, e.g., in control constructions where the control verb licenses a subject to function as a secondary subject of the controlled verb, or in relatives where the relativized noun functions as a secondary complement or adjunct within the relative clause, in addition to its external syntactic role.

2 Although HPSG analyses are directed acyclic graphs, many of the HPSG features encode trees, e.g., the DTRS feature.

2 The blurry syntax-discourse boundary and the interface problem

As noted by Carlson and Marcu (2001), the boundary between syntax and discourse is rather fuzzy, and 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."). Discourse and syntax may also interact in complicated ways. For example, long discourse units that span several sentences may function as objects of attribution verbs in direct or indirect speech, or as parenthetical remarks embedded within an otherwise normal sentence, and Wolf and Gibson (2005) and Buch-Kromann and Korzen (2010) provide examples where a complex discourse unit elaborates on a preceding NP. This raises obvious questions about how syntactic structure, which is well understood, relates to discourse structure. Since most discourse frameworks take the clause as their minimal discourse unit, there is some overlap where we can compare the intra-sentential discourse structure with the corresponding syntactic structure. When these structures differ, we must ask why they differ and how they interface with each other, given that they serve the same purpose of determining the compositional semantics and controlling the linear order, but at different linguistic levels.

It is important to note that at the intra-sentential level, discourse frameworks frequently provide structural analyses that differ from the corresponding syntactic structure, even when there is near-universal agreement about the syntactic analysis across syntactic frameworks. For example, in attributions like "The children said that they liked ice cream", the subordinate clause "that they liked ice cream" is universally analyzed as the syntactic complement of the main clause "The children said...", whereas discourse frameworks as implemented in the RST Treebank and GraphBank reverse the direction by analyzing the attribution clause as a subordinate of the attributed clause. Similarly, in discourses like "On the one hand, *X*. On the other hand, *Y*.", the two discourse adverbials "on the one hand" and "on the other hand" are universally analyzed in syntactic theories (including Lexicalized Tree-Adjoining Grammar (LTAG)) as adverbials that modify *X* and *Y*, respectively; but in D-LTAG and PDTB (Webber, 2004; Forbes-Riley et al., 2006), the two adverbials are analyzed as a single lexical item that takes *X* and *Y* as its arguments, reversing the direction of the subordination compared to syntax. In discourse (1) below, the mainstream syntactic analysis is a tree structure where "then" is analyzed as an adverbial and "when" as a subordinating conjunction; in the PDTB analysis, "then" is analyzed as the lexical anchor of an elementary tree that takes the italicized and boldfaced clauses as its argument, resulting in a completely different structure that may even be a non-tree if "when" is assumed to represent a discourse connective as well.

(1) *In an invention that drives Verdi purists bananas, Violetta lies dying in bed during the prelude, rising deliriously* **when** <u>then</u> **she remembers the great parties she used to throw**. (PDTB manual, example (36))

These differences between syntactic structure and the D-LTAG conception of discourse structure is not a problem in itself, since D-LTAG explicitly seeks to model the semantic rather than syntactic structure of discourse. But it does make it harder to reconcile PDTB's semantic conception of primary linguistic structure with the purely syntactic conception found in syntax. In the remainder of this paper, we will argue that we can reconcile the two views within a syntax-centered conception of discourse, by reframing the semantics-centered D-LTAG and PDTB conception of discourse structure as the implicit functor-argument structure associated with a single unified syntax-discourse tree structure for the entire discourse, whose elementary segments represent individual lexical items (typically words).

3 Syntax: syntactic structure vs. functor-argument structure

Syntactic theories almost universally represent syntactic structure as a tree, or a more general graph that has a primary tree as its explicit or implicit backbone, which encodes the syntactic relationships between the constituents in the sentence and constrains their linear order. The main functions of the primary tree are to control the word order and to provide an interface to semantics. Most formal semantic theories assume that phrases are assigned meanings according to the principle of compositionality, which states that the meaning of a phrase is computed as a function of the meanings of its parts (possibly supplemented with some kind of representation of the context in a dynamic semantics). We will follow Dowty (1992) and the approach taken in many linguistic theories, including HPSG, by assuming that complements are lexically selected by their governor and function as semantic arguments to their governor in the compositional semantics, whereas adjuncts lexically select their governor and function as modifiers to their governor in the compositional semantics. The intuition behind Dowty's proposal is that if we have a phrase XP with lexical head X, complement phrases $C_1,...,C_m$, and adjunct phrases $A_1,...,A_n$ (in increasing scope order), then the meaning $[XP]$ associated with the phrase is computed by first applying the functor h associated with the lexical head X of the phrase to the meanings $[C_1],...,[C_m]$ associated with the complements, and then applying the adjuncts $[A_1],...,[A_n]$ in scope-order, i.e., we define:

$$[XP] \ = \ [X + C_1...C_m + A_1...A_n]$$

which is in turn defined recursively by:

$$[X + C_1...C_m] \ = \ h([C_1], ..., [C_m])$$
$$[X + C_1...C_m + A_1...A_k] \ = \ a_k([A_k]) \, ([X + C_1...C_m + A_1...A_{k-1}]$$

where a_k denotes the functor associated with the adjunct role used to incorporate adjunct A_k into the meaning associated with XP. That is, in the syntax, the syntactic head defines the syntactic properties of the entire phrase, but semantically, each adjunct functions as a semantic head, i.e., it acts as a special kind of functor (modifier) that takes X with its complements and lower-scoped adjuncts as its argument.

Obviously, the order in which the adjuncts are applied in this meaning computation (the adjunct scope) may affect the meaning we compute for the entire phrase, i.e., two different scopes may (or may not, depending on the circumstances) lead to different meanings. Less obviously, this view of compositional semantics does not necessarily imply a conception of meaning composition as function application: it may be the case that the meaning representation associated with $[XP]$ contains the meaning representations $[C_1],...,[C_m]$ and $[A_1],...,[A_n]$ as proper substructures, but the relationship could be more complicated. For example, the meaning composition could be non-monotonic by allowing functors to change or augment substructures in the argument representations (e.g., in the treatment of free subject predicatives that act as adjuncts of the verb, although they really modify the subject from a semantic point of view). Likewise, in a dynamic semantics (cf. Groenendijk and Stokhof, 1991), the meaning composition might imply updates to the hearer's representation of the context; in such a model, expressions like parentheticals could conceivably be modelled as modifiers that affect the context exclusively without affecting the meaning of the phrase that they modify.

The distinction between phrase structure and functor-argument structure has been important in the theoretical development of syntax because it makes it possible to have two structures that serve very different purposes: a surface syntactic structure that essentially controls syntactic constraints on word order, agreement, secondary dependencies in relatives and control constructions3, etc.; and a functor-argument structure that allows for a rich and complex interface to a powerful notion of semantics, while retaining a close and well-defined interface to the syntactic structure via the notion of modifier scope. This realization did not come easily in syntax, as witnessed by the large literature on headedness in syntax (cf. Hudson, 1987; Croft, 1995; Manning, 1995).

We believe that this observation should be of interest to current theories about discourse, which do not currently seem to embody a clear distinction between syntactic and semantic structure. In their annotations, most discourse frameworks seem to lean towards a semantics-centered view where the annotations primarily encode semantic units (corresponding to the intermediate meaning representations in a functor-argument tree) and the relations between them. Today, the field seems to have moved from an initial assumption that a single tree structure may simultaneously explain the semantic interpretation and the syntactic linearization properties of discourse structure (e.g., Mann and Thompson, 1987; Polanyi, 1988; Carlson et al., 2001), to an appreciation that there do exist counter-examples where it seems difficult to find a single tree structure that reconciles these two conflicting requirements (e.g., Wolf and Gibson, 2005; Dinesh et al., 2005; Stede, 2008; Aktaş et al., 2010).

The conception of syntactic structure as a primary tree (possibly augmented with other relations) would have seemed just as untenable in syntax if syntax had been restrained to accounting for both phrase structure and functor-argument structure by means of a single tree. It therefore seems worthwhile asking whether the mechanisms

---

3   see footnote 1

that appear to have worked so well for syntax could be applied equally successfully to discourse, and what disadvantages, if any, would be associated with a shift from a semantics-centered to a syntax-centered view of discourse.

4 Discourse connectives: heads, conjunctions, markers, or adverbials?

To compare a semantics-centered and a syntax-centered conception of discourse, it is instructive to take a closer look at the analysis of discourse connectives. Discourse connectives form the backbone of the discourse annotations in the Penn Discourse Treebank and the discourse treebanks it has inspired for other languages, and seem crucial in discourse parsing: their presence as simple syntactic clues to the choice of discourse relation probably offers the best chance of getting a hold on a complex linguistic structure which is as ambiguous as it is challenging in terms of its semantic and pragmatic interpretation.

Discourse connectives are typically constructions of the form "$X\ C\ Y$", where X and Y are clauses and C is a discourse connective (such as "because", "since", "when"). Three syntactic analyses and one anaphoric analysis suggest themselves, as summarized in Table 1[4]. The analyses are drawn as dependency trees, i.e., all nodes in the tree represent elementary textual units, and the arrows go from the lexical head of a phrase to the lexical heads of its complement and adjunct phrases, with the relation name written at the arrow tip; the relation name uniquely identifies whether the dependent is a complement or adjunct. Dependency trees can be viewed as being isomorphic to restricted phrase-structure trees where every phrase has a lexical head, but depart from traditional phrase-structure trees in that discontinuous phrases (crossing branches) are allowed.[5]
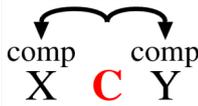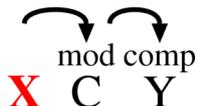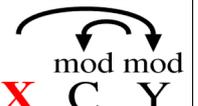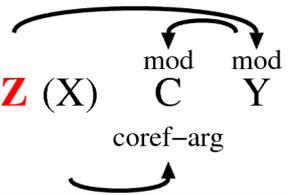
	Head	Conjunction	Marker	Anaphoric adverbial
Syntactic head	C	X	X	Z
Semantic head	C	C	Y	C
Syntax	comp comp X **C** Y	mod comp **X** C Y	mod mod **X** C Y	mod mod **Z** (X) C Y coref–arg
Semantics	C'(X',Y')	[C'(Y')] (X')	[Y'(C)] (X')	[[C'(X'')] (Y')] (Z')
CDT example	X said Y	X because Y	X and Y.	Z said X. Then Y.

Table 1. Four analyses of discourse connectives.

In the first analysis (the *head analysis*), the discourse connective C is analyzed syntactically as a head that takes X and Y as its complements; semantically, the meaning C' of C acts as functor, and the meanings X',Y' of X,Y act as arguments of C'.

---

4 The bottom row of Table 1 provides examples from the Copenhagen Dependency Treebanks (CDT)
5 Each node in the dependency tree corresponds to a possibly discontinuous phrase consisting of the yield of the node in the tree, i.e., the set of all nodes that can be reached by following the dependency arrows in the tree.

In the second analysis (the *conjunction analysis*), C is analyzed as a subordinating conjunction that takes Y as its complement and modifies X as an adjunct; semantically, C' computes a meaning $C'(Y')$ from Y', which acts as functor with X' as argument. In the third analysis (the *marker analysis*), C is analyzed as a marker that modifies Y, which in turn modifies X; semantically, Y selects a composition function $Y'(C)$ from an inventory of composition functions associated with the head of Y, and the semantically vacuous marker C merely helps disambiguate the composition function; the composition function then takes the meaning X' as its argument. In the final analysis (the *anaphoric adverbial analysis*, initially suggested by Creswell et al. (2002)), the discourse connective C retrieves its first argument X'' anaphorically, i.e., C contains an implicit anaphor X'' that provides the first argument in the discourse relation and has X as its antecedent. Syntactically, C is an adverbial that modifies Y, which in turn modifies some other discourse unit Z (we do not place any apriori restrictions on Z: it might be X itself, a unit containing X, or a completely different unit). Semantically, C is analyzed as a conjunction, i.e., C' computes a meaning $C'(Y')$ from Y', which in turn acts as functor with X' as argument; the resulting meaning $[C'(Y')](X')$ then acts as a functor which takes Z' as its argument. Note that this analysis assumes that two discourse relations are involved: one between X and Y, and another between Z and Y (possibly with a completely different connective).

The four analyses are markedly different in terms of their syntactic and semantic headedness, but similar in terms of their semantics, where X',Y' end up as arguments in all four cases (via a reference X'' to X' in the anaphoric analysis). If the discourse connective is optional, which is very often the case, the marker analysis has the obvious benefit that there is no need to postulate the presence of an implicit phonetically empty connective: the choice of composition function must then be disambiguated on the basis of semantic and pragmatic clues, rather than overt syntactic clues. 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.

Since the Penn Discourse Treebank only annotates explicit and implicit connectives, with their two arguments, the annotation itself does not specify which of the four syntactic analyses defined above applies to the individual annotations. But from the work on D-LTAG (Forbes-Riley et al., 2006), the theoretical framework that informs the annotation of the Penn Discourse Treebank, it appears that D-LTAG analyzes subordinating conjunctions like "although" as initial trees (essentially a head analysis), coordinating conjunctions like "and" and "but" are analyzed as auxiliary trees (essentially a conjunction analysis, with a phonetically empty connective if the connective is implicit), discourse adverbials like "then" are analyzed as discourse adverbials, and parallel adverbial constructions like "On the one hand, X. On the other hand, Y" are analyzed as initial trees (head analysis). Interestingly, although D-LTAG is based on the syntactic framework LTAG, D-LTAG differs from LTAG in its analysis of subordinating conjunctions and parallel adverbial clauses: D-LTAG uses a head analysis for these constructions, instead of the conjunction analysis an

adjunct analysis used in LTAG and most other syntactic frameworks.[6]

5 Attribution: a difficult case requiring the full power of compositional semantics

As pointed out by Dinesh et al. (2005), attribution is one of the main obstacles for a syntax-centered conception of discourse. Consider their discourse analysis in (2) below:

(2)　*The current distribution arrangement ends in March 1990*, <u>although</u> 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.]

Ignoring the final attribution to the spokeswoman, the discourse is of the form "*X* although Delmed said *Y*". The problem here is that mainstream syntax universally analyzes "Delmed said *Y*" as the complement of "although", but the most sensible reading of (2) is that the discourse relation signalled by "although" holds between *X* and *Y*, rather than between *X* and Delmed's saying event. Carlson et al. (2001) and Wolf and Gibson (2005) try to circumvent this problem by analyzing the attribution as a satellite and the attributed event as the nucleus, but this does not really solve the problem, since the discourse relation may also refer to the attribution event, as demonstrated by (3):

(3)　*Advocates said the 90-cent-an-hour rise, to $4.25 an hour by April 1991, is too small for the working poor,* <u>while</u> **opponents argued that the increase will still hurt small business and cost many thousands of jobs**. [(13) in Dinesh et al.]

Dinesh et al. suggest that the problems with attribution could be taken as arguments against a tree-structured discourse, which would undermine a syntax-based view of discourse. We would like to propose two alternative responses – the first accepts the analysis of these examples given by Dinesh et al., while the second proposal relies on a different analysis.

　　Our first proposal involves the introduction of a more powerful compositional mechanism to address the problem pointed to by Dinesh et al. Given the highly complex compositional semantic mechanisms that are needed in syntax, in any case (e.g., for markers and Pustejovsky-style lexical semantics), we find this is a reasonable response, rather than giving up the idea that discourse structure can be modelled by a syntactic tree.

　　Specifically, suppose we have a discourse of the form "*X C Y*" where *X* and *Y* may contain a chain of attributions (i.e., *Y* could be of the form "Delmed said *Z*", "Delmed said Ann claimed *Z*", "Delmed said Ann claimed Bob believed *Z*", etc.). Let *c* denote the standard composition function associated with *C*, and suppose π is an operator

---

6　Since the purpose of D-LTAG is to perform discourse parsing, it is quite possible that this change in analysis is motivated by computational rather than linguistic considerations.

that given an epistemic formula $K_a\varphi$ ("φ is known by agent a") returns φ. In order to handle attributions in the compositional semantics, we only have to assume that instead of letting C have a single composition function c which given arguments X',Y' computes a meaning representation $c(X',Y')$, it has a whole family of composition functions c_{ij} defined by $c_{ij}(X',Y') = c(\pi^i(X'),\pi^j(Y'))$ where i,j cannot exceed the length of the attribution chain in X,Y. When computing the compositional semantics, we then have to disambiguate not only the correct relation associated with C, but also the correct choice of i,j.

This step is not as radical as it may seem at first sight. Many explicit discourse connectives seem to support more than one reading, i.e., they have more than one natural composition function. If we also adopt the marker analysis, we are in principle assuming that any discourse unit can attach to any other discourse unit, choosing a composition function from the full inventory of discourse relations on the basis of contextual clues and optional syntactic clues. In this case, our compositional treatment of attribution essentially just means adding a little more ambiguity to the set of composition functions provided by the inventory of possibly implicit discourse relations.

The compositional account of attribution does not prevent us from making a precise annotation either, since we can disambiguate the correct choice of numbers i,j for a relation R by annotating the relation as "iRj" rather than "R" – this is actually the essence of the annotation scheme for attribution used in the Copenhagen Dependency Treebanks (Buch-Kromann and Korzen, 2010), except that i and j are annotated as sequences of asterisks, rather than as numbers. Attribution is therefore not as big an obstacle to a syntax-centered conception of discourse that it might at first appear to be.

Our second response calls into question the analysis given of example (2) by Dinesh et al. – the key problem is that *although* relates X with Y, rather than relating X with "*Delmed said Y*". The syntax-discourse mismatch is eliminated if it is possible to analyze "*Delmed said Y*" as the second argument of the contrast relation, and we argue that this indeed is the proper analysis here. In fact, it is typical for contrastive relations to arise between conflicting propositions from different sources: in fact that is precisely the situation in example (3), as Dinesh et al. point out. The only difference in (2) is that the first argument is *implicitly* associated with the speaker, while the second argument is explicitly associated with Delmed. In our view, it is quite natural to contrast the two under the assumption that Delmed is credible.

It may well be that there are cases of attribution that require an analysis that reveals a mismatch between syntax and discourse. But in our view, examples (2) and (3) from Dinesh et al. are properly analyzed without any such mismatch. Thus while we are open to the possibility that the more complex compositional mechanism may indeed be necessary, we leave the issue unresolved in this paper.

6 Tree structured discourse: the counter examples from a syntax-centered view

A lot of research in discourse structure has centered on the question whether discourse structure can be viewed as a tree structure or not. Wolf and Gibson (2005)

were among the first to question the suitability of tree structures for discourse, followed by many other researchers, including Dinesh et al. (2005), Lee et al. (2006, 2008), Stede (2008), and Aktaş et al. (2010).

Wolf and Gibson (2005) created a corpus of discourse analyses, without requiring the analyses to be trees, and found that the resulting analyses deviated significantly from trees by including crossing relations and multi-nuclearity. In a syntax-centered conception of discourse, Wolf and Gibson's finding with respect to crossing relations only shows that discourse resembles syntax in this respect, since discontinuous word order phenomena are a key issue in syntactic frameworks, and all sophisticated syntactic theories have a complex set of mechanisms to account for this challenge. Multi-nuclearity is much harder to reconcile with a syntax-centered view of discourse, but here we essentially agree with the counter-criticism voiced by Marcu (2003), who argued that some of the additional relations were really coreference relations, and the remaining counter-examples might be an artefact of their annotation conventions; this view is mostly supported by Knott (2007).

Dinesh et al. (2005) compared the annotations of subordinating conjunctions in the Penn Discourse Treebank (PDTB) with the syntactic annotations in the Penn Treebank (PTB). They found that there were significant differences between the analysis of syntax and discourse, most of which were caused by the treatment of attribution in the PDTB. The problems associated with attribution was addressed in the preceding section, and we believe some of their other counter-examples can be explained by other means: in some cases, the analysis is ambiguous in both the syntactic annotation and the discourse annotation, and the PTB annotators did not choose the same analysis as the PDTB annotators (e.g., their examples (14)-(15)); in other cases, the analysis chosen by PDTB could have been obtained by assuming a particular modifier scope in the syntactic analysis (e.g., their examples (16)-(17)); differences may also be caused by the coarser granularity of the segmentation in the PDTB (e.g., their examples (18)-(19)).

Lee et al. (2006) provide additional examples of complex discourses from the PDTB that violate one or more tree constraints, including examples of independent relations, shared arguments, properly contained arguments, pure crossings, and partially overlapping arguments. We will follow their formatting conventions, using boldface for the arguments of the first connective, and italics for the arguments of the second connective. Their example of shared arguments has the form "**X** but *Y* so *Z*", which looks very different from a syntax tree, especially because they seem to draw the functor-argument tree rather than the syntax tree (if it was a syntax tree, they would be using a head analysis). In the mainstream syntactic analysis of this example, "so Z" modifies "Y" and "but Y so Z" modifies "X", with the connectives analyzed as either conjunctions or markers, depending on personal preference (cf. section 4). This would give the functor-argument structure "**X** but *Y* so *Z*", i.e., we can obtain the same semantic analysis as in the PDTB if we allow the compositional semantics of "but" to strip off "so Z" from "Y so Z" before the composition with "X", a strategy that does not seem to be completely untenable, given the complexity of the compositional semantics in many other respects. We believe that this

mechanism, coupled with the anaphoric discourse adverbial analysis proposed by Forbes-Riley et al. (2006), can explain the examples of properly contained arguments, pure crossings, and partially overlapping arguments given by Lee et al. In their example of independent relations, consisting of two unconnected trees, the second tree could be analyzed as an elaboration of an NP in the first tree. The examples provided by Aktaş et al. (2010) follow essentially the same pattern as in Lee et al., and we believe they can be accounted for by means of the same mechanisms.

Stede (2008) considers a range of criteria that could be used to determine the analysis of nuclearity in a discourse, including the intention of the text (which segment is most central to the writer's purposes), the thematic development of the text (recurrence, repetition, digression meta-discursive element), surface-oriented properties (connectives, other lexical marking, syntactic structure), and specific conventions adopted by the annotation scheme. He argues that these criteria are often conflicting, in particular, that it is possible to find examples where the writer's purposes run against the syntactic subordination. These counter-examples are typically of the form "X Y Z", where X and Y are related by a multi-nuclear relation, i.e., either of them could function as the nucleus, and Z can be manipulated so that it is a satellite of either X or Y. Stede's argument is that in a discourse structure based on trees where crossing relations are disallowed, we will be forced to select different analyses of the relationship between X and Y, i.e., the nucleus of "X Y" necessarily coincides with the nucleus for Z. However, these examples could just as well be taken as evidence for crossing relations, which would not be problematic in a syntax-based conception of discourse. In other cases where Stede departs from the syntactic analysis in his discourse analysis, he does so because he sees the syntactic structure as peripheral to Mann and Thompson's characterization of the nucleus as being "more central to the writer's purposes", i.e., he essentially reverses the syntactic relation in order to ensure that the unit which would be most important in a summary of the text is selected as the nucleus. However, this could equally well be seen as an argument for placing less emphasis on centrality and more emphasis on syntactic structure. In fact, Stede acknowledges that what is central to one's purposes can be very different from case to case, an observation that points to an important weakness in the notion of centrality. Incidentally, the notion of semantic prominence has been given up as a main criterion for headedness in syntax: syntactic theories routinely assume that main verbs may function as complements of auxiliaries and modals, although semantic prominence would seem to argue for the converse analysis.

In conclusion, it seems that many of the examples that have been suggested as counter-evidence against a tree-based discourse analysis that spans the entire discourse, can be accommodated within a syntax-centered discourse framework with a flexible compositional semantics and a rich set of mechanisms, e.g., for dealing with anaphoric discourse adverbials. In our view, the most intriguing outcome of this discussion is Stede's observation that the syntactic structure sometimes differs significantly from what is central to the writer's purposes. We could draw the conclusion that syntactic structure is less important than centrality, but the reverse

conclusion is just as possible: that it is the syntactic structure which is the more important of the two, which seems to be the near-universal conclusion in syntax.

7 Ambiguity and other remaining problems in the syntax-based view

A syntax-centered discourse annotation solves a number of problems – in particular, it allows the discourse to be represented as a tree with additional relations for coreference. But it also introduces some problems as well. Most importantly, whenever a nucleus has more than one adjunct, we can only compute the functor-argument structure if we are given a modifier scope. More generally, if we impose a highly principled linguistic framework on our annotation of discourse, including a tree-based model, it is obviously difficult to use these data to argue for the particular assumptions. This is however not a crucial objection, since exactly the same thing could be said about syntactic annotation.

8 The syntax-based discourse annotation in the CDT treebanks

The kind of syntax-centered discourse annotation we have described in the paper is currently being implemented in the Copenhagen Dependency Treebanks (CDT) to create a set of open-source parallel treebanks for five different languages, Danish, English, German, Italian, and Spanish (cf. Buch-Kromann et al., 2009; Buch-Kromann and Korzen, 2010). These treebanks resemble the Potsdam Commentary Corpus (Stede 2008) in that they are multi-level treebanks, i.e., the annotation includes syntactic structure, discourse structure, and coreference structure. The annotation is in its early stages, but more than 273 text excerpts with approximately 250 words in each excerpt have been annotated so far, using a detailed inventory of 50 discourse relations organized in a hierarchy so that different levels of granularity can be selected. The current inter-annotator agreement is approximately 50%, a number that we hope to improve.

9 Conclusions

The important question that we have sought to answer in this paper is whether discourse structure and syntactic structure are fundamentally different structures, or whether they are better viewed as instances of a single unified syntax-discourse tree structure at different levels of granularity in the segmentation. We have argued that if we think of discourse structure as an extended syntactic structure with an induced, but not explicitly expressed semantic predicate-argument structure that links the sentences in the entire discourse, the second, unified view is not only feasible, it also solves a number of syntax-discourse interface problems and provides a unified view of syntax and discourse that should make it easier to extend almost-linear parsing algorithms like the Malt parser (Nivre, 2006) to discourse parsing.

We have also argued that a syntax-centered view of discourse involves a significant departure from the original definition of nuclearity in Rhetorical Structure Theory, which is based on the notion of centrality to the writer's purposes, to a much more surface-oriented view of discourse structure. One possible concern is that discourse-based tasks like text summarization may become much harder in a syntax-

centered conception of discourse; on the other hand, discourse parsing might become easier because the resulting analyses are closer to the syntactic analyses.

References

Berfin Aktaş, Cem Bozşahin, Deniz Zeyrek, 2010. *Discourse Relation Configurations in Turkish and an Annotation Environment.* Proc. Linguistic Annotation Workshop (ACL-2010).

Matthias Buch-Kromann. 2009. *Discontinuous Grammar. A dependency-based model of human parsing and language learning.* VDM Verlag.

Matthias Buch-Kromann, Iørn Korzen & Henrik Høeg Müller. 2009. Uncovering the 'lost' structure of translations with parallel treebanks. *Copenhagen Studies in Language* 38: 199-224.

Matthias Buch-Kromann and Iørn Korzen. 2010. The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks. Proc. Linguistic Annotation Workshop (ACL-2010).

Lynn Carlson and Daniel Marcu. 2001. *Discourse Tagging Reference Manual.* ISI Technical Report ISI-TR-545.

Lynn Carlson, Daniel Marcu, Mary Ellen Okurowski. 2001. *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory.* Proc. Discourse and Dialogue.

Noam Chomsky. 1981/1993. *Lectures on Government and Binding: The Pisa Lectures.* Mouton de Gruyter.

Cassandre Creswell, Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Bonnie Webber, Aravind Joshi. 2002. *The discourse anaphoric properties of connectives.* Proc. DAARC 2002, pages 45-50.

William Croft. 1993. *What is a head*? In J. Rooryck and L. Zarin (eds.), *Phrase structure and the lexicon*, pages 35–76. Dordrecht: Kluwer Academic Publishers.

Mary Dalrymple, Ronald M. Kaplan, John Maxwell III, Annie Zaenen (eds.). 1994. *Formal issues in Lexical-Functional Grammar.* CSLI Lecture Notes, no. 47.

Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, 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,* pages 29-36.

David R. Dowty. 1992. Towards a minimalist theory of syntactic structure. In: Wietske Sijtsma and Arthur van Horck (eds.), *Discontinuous constituency*, Mouton de Gruyter.

Denys Duchier. 2001. Topological dependency trees: A constraint-based account of linear precedence. Proc. ACL-2001.

Katherine Forbes-Riley, Bonnie Webber, Aravind Joshi. 2006. Computing Discourse Semantics: The Predicate-Argument Semantics of Discourse Connectives in D-LTAG. *Journal of Semantics*, 23(1).

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy* 14:39-100.

Richard Hudson. 1987. Zwicky on heads. *Journal of Linguistics* (23): 109-132.

Richard Hudson. 2010. *An introduction to Word Grammar.* Cambridge University Press.

Aravind Joshi and Yves Schabes. 1997. Tree-adjoining grammars. In: Grzegorz Rozenberg and Arto Salomaa (eds.), *Handbook of Formal Languages. Beyond Words.* Springer-Verlag.

Alistair Knott. 2007. Review of 'Coherence in natural language: Data structures and applications', by Florian Wolf and Edward Gibson. *Computational Linguistics* 33:591–595.

Alex Lascarides and Nicholas Asher. 2007. Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure. In: Harry Bunt and Reinhard Muskens (ed.), *Computing Meaning.* Synthese language library, vol. 83. Springer Netherlands, pages 87-124.

Alan Lee, Rashmi Prasad, Aravind Joshi, Bonnie Webber. 2008. *Departures from Tree Structures in Discourse: Shared Arguments in the Penn Discourse Treebank.* Proc. Constraints in Discourse III Workshop.

Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Bonnie Webber. 2006. *Complexity of Dependencies in Discourse: Are Dependencies in Discourse More Complex Than in Syntax?* Proc. Treebanks and Linguistic Theories. Prague, Czech Republic. December 2006

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.

Daniel Marcu. 2003. Discourse structures: trees or graphs http://www.isi.edu/~marcu/discourse/Discourse%20structures.htm

Igor Mel'čuk. 1988. *Dependency syntax*. State University of New York Press.

Lucie Mladová, Šarka Zikánová, Eva Hajičová. 2008. *From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank.* Proc. LREC-2008.

Joakim Nivre. 2006. *Inductive Dependency Parsing.* Springer.

Livia Polanyi. 1988. A Formal Model of Discourse Structure. Journal of Pragmatics 12: 601-639.

Carl Pollard and Ivan A. Sag. 1994. *Head-driven phrase structure grammar*. Univ. of Chicago Press.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proc. 6th Int. Conf. on Language Resources and Evaluation*, Marrakech, Morocco.

Petr Sgall, Eva Hajičová, Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: D. Reidel Publishing Company.

Manfred Stede. 2008. RST revisited: Disentangling nuclearity. In: C. Fabricius-Hansen, W. Ramm (Eds.): *'Subordination' versus 'coordination' in sentence and text - A cross-linguistic perspective.* Studies in Language Companion Series. Amsterdam: John Benjamins.

Mark Steedman. 2000. The syntactic process. A Bradford Book, The MIT Press.

Bonnie Webber. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science* 28: 751-779.

Florian Wolf and Edward Gibson 2005. Representing Discourse Coherence: A Corpus-Based Study. Computational Linguistics 31(2), 249-287.