D-LTAG System: Discourse Parsing with a Lexicalized Tree Adjoining Grammar *

Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad
(forbesk,elenimi,rjprasad@linc.cis.upenn.edu)
Institute for Research in Cognitive Science, University of Pennsylvania, 3401 Walnut Street, Philadelphia, PA 19104, USA

Anoop Sarkar (anoop@cs.sfu.ca)
School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada

Aravind Joshi (joshi@linc.cis.upenn.edu)
Institute for Research in Cognitive Science, University of Pennsylvania, 3401 Walnut Street, Philadelphia, PA 19104, USA

Bonnie Webber (bonnie@inf.ed.ac.uk)
Division of Informatics, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland

Abstract. We present an implementation of a discourse parsing system for a lexicalized Tree-Ajoining Grammar for discourse, specifying the integration of sentence and discourse level processing. Our system is based on the assumption that the compositional aspects of semantics at the discourse-level parallel those at the sentence-level. This coupling is achieved by factoring away inferential semantics and anaphoric features of discourse connectives. Computationally, this parallelism is achieved because both the sentence and discourse grammar are LTAG-based and the same parser works at both levels. The approach to an LTAG for discourse has been developed by Webber et al. in some recent papers ([33], [35], among others). Our system takes a discourse as input, parses the sentences individually, extracts the basic discourse consituent units from the sentence derivations, and reparses the discourse with reference to the discourse grammar while using the same parser used at the sentence-level.

1. Introduction

All work on discourse starts from the premise that discourse meaning is more than the sum of its parts (i.e., its constituent sentences or clauses). The question is how to get there. Work in the tradition of *Rhetorical Structure Theory* [17] - both in interpretation [18] and generation [8] - views the additional meaning solely in terms of discourse relations that hold between adjacent text spans, treating discourse connectives as signalling types of discourse relations. How the basic text spans are as-

* The first four authors are named in alphabetical order.

signed an interpretation, and how that interpretation might contribute to discourse meaning apart from discourse relations, is largely ignored.

Not so in more formal work on discourse semantics ([4], [21], [25], [26], among others), which takes seriously a compositional process of deriving discourse meaning from the meaning of its parts. However, this work (a) only makes use of two mechanisms for deriving discourse meaning from the meaning of its parts - compositional semantics and inference - and (b) treats the process by which discourse derives compositional aspects of meaning as being entirely separate from how clauses do so. Both of these are the focus of the approach developed in [33] and [35]. In this approach, it is argued that certain aspects of discourse meaning are better seen as deriving from anaphoric and presuppositional properties of lexical items, and that this is facilitated through a uniform lexicalised treatment of both clausal syntax and semantics as well as discourse syntax and semantics. This paper presents an implementation of a discourse parsing system that draws on the insights of this latter approach. We explore the hypothesis that the boundary between sentence level structure and discourse level structure is not a sharp one. Sentence level structure supports compositional semantics even though there are other semantic aspects, such as anaphoric relations (e.g., intrasentential links for pronoun reference) and inferential interpretation (e.g., interpretation of compound nouns) that need to be accounted for. In the same way, discourse level structure is also viewed as supporting compositional aspects of semantics, while allowing for other interpretive components to be added on for a complete semantics for discourse - e.g., for determining anaphoric and inferential interpretation. Thus, we pursue the idea that the formal device used for deriving the structural descriptions at both levels is the same, while noting that at the discourse level, the device may have less generative power. In addition, we also illustrate that our architecture for discourse parsing allows for a smooth transition from sentence level to discourse level processing and for the use of a single parser at both levels.

In Section 2, we discuss the framework for an LTAG for discourse. Section 3 presents a discussion of our methodology for determining the structure of discourse connectives, accompanied with a case study of the connective *however*. In Section 4, we describe the architecture of our system. Section 5 discusses some of the advantages of our system, in particular, with respect to the close link between sentence level and discourse level semantics, and with respect to the projection of sentence-level information structure distinctions to the discourse level. Finally, in Section 6, we compare our system with some other approaches, especially with those that use some variant of TAG for describing discourse structure (e.g., [4] and [26]), and those that attempt to automate the derivation of discourse structure (e.g., [18]).

2. A Lexicalized Tree Adjoining Grammar for Discourse

The D-LTAG system is based on the approach to a lexicalized TAG for discourse, as described in [33]. An LTAG for discourse posits two kinds of elementary trees: *initial* trees, which encode predicate-argument dependencies, and *auxiliary* trees, which are recursive and modify and/or elaborate elementary trees. All structural composition is achieved with two operations, *substitution* and *adjunction*. In the D-LTAG grammr, clauses connected by a subordinating conjunction form an initial tree whose compositional semantics is determined by the semantic requirements of the subordinate conjunction (the predicate) on its arguments (the clauses). Auxiliary trees are used for providing further information through adjunction. They can be anchored by adverbials, by conjunctions like *and*, or may have no lexical realization. Furthermore, a discourse predicate may take all its arguments structurally, as in the case of subordinating conjunctions, or anaphorically, by making use of events or situations available from the previous discourse, as in the case of *then*.[1] This division between the compositional part of discourse meaning (projected by the tree structures) and the non-compositional contributions due to general inferencing and anaphora is a key insight of the approach to an LTAG for discourse. It simplifies the structure of discourse and extends compositional semantic representations from the sentence level to the discourse.

Figure (1a) shows one initial tree in the grammar.[2] We treat connectives anchoring this tree as discourse predicates which require two clausal arguments. In general, such trees are anchored by subordinating conjunctions such as *because*, *when* etc. A second initial tree is shown in Figure (1b) which, as suggested in previous work [33], reflects dependencies in parallel constructions and is projected by pairs of connectives such as *on the one hand ... on the other hand...*[3]

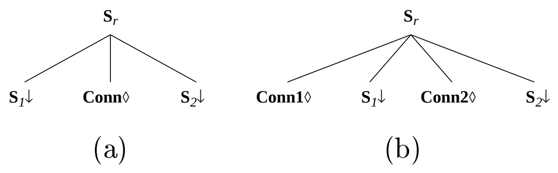

Figure 1. Initial Trees in D-LTAG

As in the sentence-level grammar, trees in the D-LTAG grammar are grouped into *tree families*, which are taken to reflect surface clause order variation realized with preposed and postposed subordinate clauses. Trees belonging to the same family share the same predicate-argument dependencies. One such tree family is shown in Figure 2, anchored by connectives like *because*.

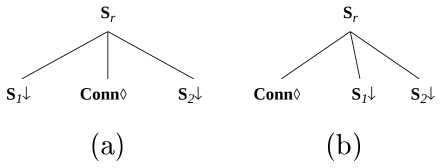

Figure 2. Tree Family in D-LTAG

The second type of elementary tree consists of *auxiliary trees*, which introduce recursion and serve to extend or modify a description of the previous discourse. There are two kinds of auxiliary trees, shown in Figure 3.

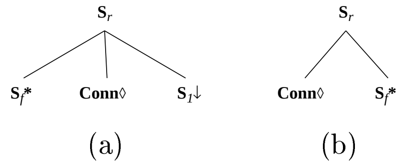

Figure 3. Auxiliary Trees in D-LTAG

The tree in Figure 3(a) is anchored by connectives that simply continue the description conveyed by the structure to which the tree adjoins. Other aspects of meaning that relate the two arguments are derived anaphorically or inferentially (e.g., based on the relationship between the tense/aspect of the two arguments ([7], [12], [16], [32]). The anchor of this tree can also remain lexically unrealized, when it is used to connect adjacent clauses without overt connectives, such as "*Mary walked towards the car. The door was open*". The tree in Figure 3(b) is selected by connectives whose first argument is resolved anaphorically and the second argument is the interpretation of the clause they adjoin to. We say more about this in the next section.

3. Determining Tree Structures for Discourse Connectives

In the previous section, we defined the elementary trees included in the D-LTAG grammar. The next crucial step in lexicalizing a Tree-Adjoining Grammar is determining which trees or family of trees are selected by a discourse connective. In previous work ([35], [34]), it was shown that the connectives *then*, *for example* and *otherwise* are best treated as anaphoric, anchoring trees of type 3(b). But for some other connectives, such as *however*, it was less clear whether they were structural or anaphoric.

In what follows, we take the view that the lexicalization of trees is an empirical question, and we describe the methodology we adopt to determine the structures lexicalized by connectives. When in doubt about the structure of a certain connective, we start with the hypothesis that the arguments of the connective are realized structurally. This is because, from a computational point of view, it would be less interesting to start with the assumption that arguments are resolved anaphorically. Assuming that all connectives find their arguments anaphorically would probably be adequate to characterize all predicate-argument relationships on the discourse level. However, it would not shed much light on those aspects of structural organization that are relevant to language structure and presumably contribute to the efficiency of the inferencing processes required for discourse interpretation [10]. Predicates which find their arguments structurally define a domain of locality for structural dependencies and constrain the interpretation of discourse in a computationally efficient way, as is the case for verb predicates at the sentence level syntax.

On empirical grounds, the diagnostic we use to test if a connective is structural is crossed structural dependencies. The current XTAG grammar for English does not lead to crossed dependencies as they seem to be unnecessary at the sentence level for English. We make a similar assumption for the discourse level and conclude that a connective defines a domain of structural locality only when such domains do not cross tree nodes.[4]

A Case Study - "However": For the connective *however*, our first assumption is that *however* anchors the structural auxiliary tree, shown in Figure 3(a). Regarding its semantic contribution, we follow [13] and [15] in assuming that *however* presupposes a defeasible rule holding between a generalization of its first argument and a generalization of the negation of its second argument, and asserts that the rule fails to hold in this case (see [36] for a formalization of the rule). To investigate if both arguments are realized structurally, we conducted a corpus study of the connective. We identified 71 tokens of *however* from the Brown corpus and located the two arguments of the connective for each token. In 69 out of the 71 instances, the position of both arguments in the discourse was consistent with the structural hypothesis: one argument was realized in the sentence containing the connective, and the second argument was realized either in the immediately preceding sentence

(58 instances) or in an immediately preceding chunk of text (11 instances). In both cases, attachment to the previous discourse did not yield crossing of tree nodes.

The remaining 2 cases were of two kinds. One, exemplified in (1), involved an argument that was not directly realized in the previous discourse. Rather, the presupposed defeasible rule could only be seen as holding between rather complicated generalizations which would have to be *inferred* from the two arguments. Here we take the defeasible rule to be something like "If the speaker/writer makes an apparently negative comment about a book, then his/her attitude is negative towards it."

(1) a. If this new Bible does not increase in significance by repeated readings throughout the years, it will not survive the ages as has the King James Version.

b. *However*, an initial perusal and comparison of some of the famous passages with the same parts of other versions seems to speak well of the efforts of the British Biblical scholars.

In the other case, *however* appeared to make no semantic contribution to the discourse, other than simple continuation. This is shown in (2).

(2) a. It is in this spirit which explains some of the anomalies of American Catholic higher education, in particular the wasteful duplication apparent in some areas.

b. I think for example of three women's colleges with pitiful enrollments, clustered within a few miles of a major Catholic university, which is also co-educational.

c. This is not an isolated example;

d. this aspect of the total picture has been commented upon often enough.

e. It would seem to represent esprit de corps run riot.

f. Apart, *however*, from the question of wasteful duplication, there is another aspect of the "family business" spirit of Catholic higher education that deserves closer scrutiny.

While it is clear that (2f) attaches higher up to the structure containing (2b)-(2e), it is less clear what the semantic contribution of *however* is to the interpretation of the discourse. *However* here seems to be acting similar to the discourse marker *now* (e.g., "Now, apart from the question of wasteful duplication...") [6], reinforcing the IRU (Informationally Redundant Utterances) cue (i.e., "apart from the question of wasteful duplication") as a signal of returning to (2a) after a conceptually embedded segment was closed off at (2e) ([5], [31], [3]).

To summarize, the corpus-based study for the connective *however* provides considerable support for the hypothesis that it finds its arguments structurally. However, as indicated by the more complex examples (1) and (2), further empirical studies will be required before a definitive conclusion is reached.[5]

4. System Description and Implementation

In this section, we describe our initial implementation of a discourse parsing system based on a lexicalized Tree-Adjoining Grammar for discourse. Discourse structure is derived in two passes of parsing. In the first pass, the sentences in the discourse are parsed, whereas discourse parsing is done in the second pass. Without losing sight of the key ideas of the theory of an LTAG for discourse, this two pass implementation achieves a considerable simplification over a single pass of parsing, especially in terms of the parsing time and space requirements that would result from using both the sentence-level and the discourse-level grammar at once.

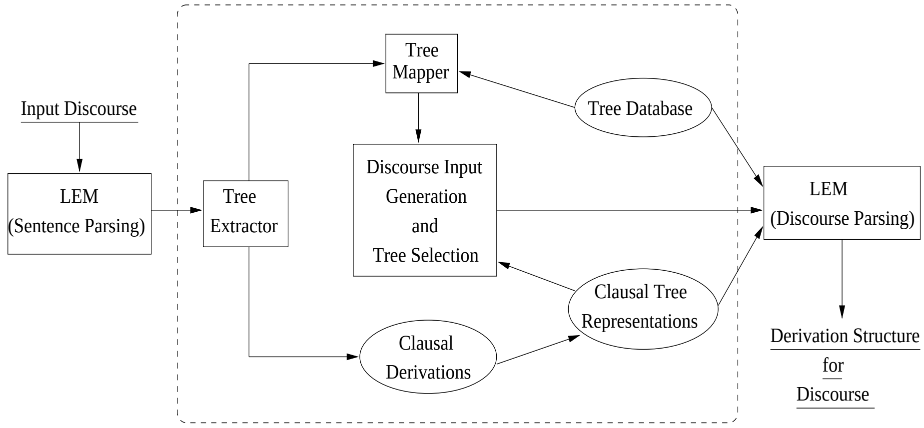

Figure 4. D-LTAG: System Description

Figure 4 shows the overall architecture of the system. The input discourse is submitted to LEM, the **Parser**, which parses each sentence in the discourse with reference to the sentence grammar. The output derivations (one derivation each for each of the sentences) are then submitted to the **Tree Extractor**, which extracts the basic discourse constituent units from each sentence derivation. The basic discourse units constitute the elementary trees lexicalized by discourse connectives in the sentence-level grammar, and the derivation (and derived) structures associated with the clausal units.[6] In the next step, the sentence-level elementary trees anchored by the connectives are mapped by the **Tree Mapper** to their corresponding elementary trees in the discourse grammar. The discourse grammar, as specified in Section 2, is contained in the **Tree Database**. The output of the Tree Mapper, together with the clausal units and the input discourse string, is then used to construct a **discourse input representation** that consists of a sequence of lexicalized trees (tree selection), with the extracted connectives and clausal units as the lexicalizing elements.[7] Finally, the discourse input, the Tree Database, and the clausal tree representations are submitted to the same **Parser**, which provides derivations for the given discourse.

In the rest of this section, we describe the different components of the system in greater detail, and discuss various issues that arose during the implementation.

PARSER (LEM).The parser is a chart-based head-corner parser [23]. The sentence-level grammar used by the parser is the XTAG English grammar [39]. For each sentence, each subsequent phase of the system assumes that there is exactly one derivation per sentence. Since, in general, there can be many ambiguities for each sentence in the discourse, the parser picks one derivation per sentence to pass on to subsequent processing. In the system described in this paper, the parser produces a single parse for each sentence by using heuristics that (a) decide which elementary tree to assign to each word, and (b) pick the lowest attachment between these trees. In future work, we plan to experiment with two other methods to deal with ambiguity: (1) using the parser as a statistical parser [24] where it reports the most probable parse based on training the parser on the Penn Treebank, and (2) representing the many parses associated with each clausal unit in the sentence in a compact form (a parse forest) and representing these as the elementary units in the discourse.

TREE EXTRACTOR. The task of this component is to extract, from each sentence derivation, the clausal derivations and any elementary tree(s) anchored by discourse connectives. The Extractor first does a top-down traversal of the sentence derivation, and identifies the part of the derivation associated with any connectives. Identification of the connectives is done against a database containing a list of possible discourse connectives as well as the elementary tree(s) anchored by each of them in the sentence grammar.

The use of both lexical and structural information is necessary to correctly identify the discourse usages of connectives in the sentential derivations. That is, neither kind of information by itself is sufficient for identification. On the one hand, many elements that function as discourse connectives can also have other functions: *and* functions as a discourse connective when it conjoins clauses, as in "The dog barked *and* Mary smiled", but it can also conjoin noun phrases (among other phrasal categories), as in "Lana ate cheese *and* crackers". As a result, if we used only the lexical appearance of the elements as the identification criterion, then the *and* which conjoins non-sentential categories would be incorrectly treated as a discourse connective. Knowledge about the elementary trees associated with the discourse usage of *and* is therefore necessary to rule this out. On the other hand, it is not sufficient to only use structural information to identify discourse connectives. For example, the sentence-level grammar does not make a structural distinction between sentential adverbs that *are* discourse connectives and those that are *not*: the elementary tree in Figure 5 can be lexicalized both by *however*, which *is* a connective, as well as by *always*, which is not a connective. Identification of the former - and not the latter - structure can therefore be done only if the lexicalizing elements are also used.

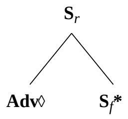

Figure 5. Elementary Tree Anchored by an Adverb in the Sentence-level Grammar

After the identification of the connectives, the clausal derivations are detached in the sentence derivation at the substitution and/or adjunction nodes of the connective elementary tree. The result of this procedure is shown in Figure 6 for the derivation of the sentence *while she was eating lunch, she saw a dog.*[8]

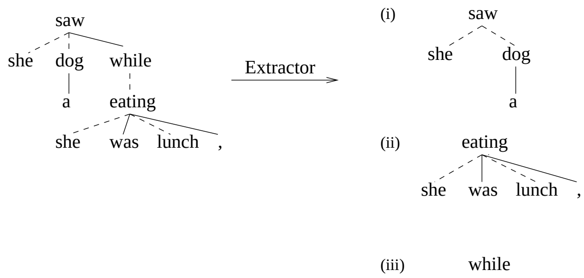

Figure 6. Result of Tree Extractor Applied to Derivation of *while she was eating lunch, she saw a dog.*

The above procedure can be shown to work on all derivations in which connectives take clauses as their arguments in the sentence-level grammar. In the surface string, this corresponds to connectives appearing at clause boundaries. However, connectives can also appear in clause-medial positions, as in Example 3. The connective *then* adjoins to the verb phrase (VP node) in the clause.

(3) Susan will *then* take dancing lessons.

Though such clause-medial connectives are posited as taking clauses as their arguments in the discourse-level grammar, we believe that their clause-internal syntax should be visible at the discourse-level description, as it is probably an indicator of Information Structure (IS).[9] The Extractor achieves this by only making a *copy* of the derivations for these connectives, and by replacing - in the derivation of the clause - the lexical occurrence of the connective by an index, to indicate its clause-internal position. The result of this procedure for example (3) is given in Figure 7. {*then*} in 7(i) represents the clause-medial connective index left by the Extractor.

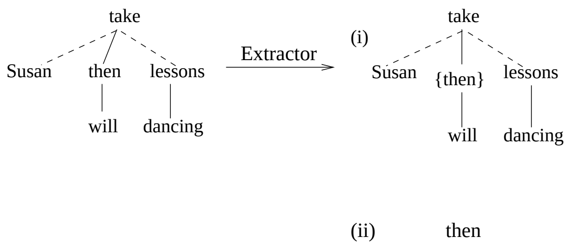

Figure 7. Result of Tree Extractor Applied to Derivation of *Susan will then take dancing lessons*

Thus, the output of the Extractor is, for each sentence, a set of elementary trees anchored by connectives in the sentence grammar, and a set of clausal derivations. For the example discourse given in (4), (5) shows the parts of the discourse input corresponding to the extracted discourse constituent units. (b), (f), (i) and (k) are the extracted connectives, and the rest are the clausal derivations.

(4) a. Mary was amazed.
b. While she was eating lunch, she saw a dog.
c. She'd seen a lot of dogs, but this dog was amazing.
d. The dog barked and Mary smiled.
e. Then she gave it a sandwich.

(5) a. mary was amazed
b. while
c. she was eating lunch
d. she saw a dog
e. she'd seen a lot of dogs
f. but

g. this dog was amazing
h. the dog barked
i. and
j. mary smiled
k. then
l. {Then} she gave it a sandwich

TREE MAPPER. The connective-lexicalized elementary trees that are extracted from the sentence derivations are submitted to the Tree mapper, which maps their sentence-level structural descriptions to their discourse-level structural descriptions (taken from the Tree Database). This is a crucial step in the discourse derivation because it is involved with determining what kinds of contribution(s) a given connective makes to discourse meaning, that is, what it contributes through compositional semantics, through anaphora and through inference. Furthermore, as has been pointed out in Section 3, determining the discourse structures anchored by connectives is an empirical matter. A major part of the future work in this project is to fully determine this mapping with corpus based work on the behavior of connectives. We continue here by assuming the mappings shown in Figure 8 for the example discourse (4).

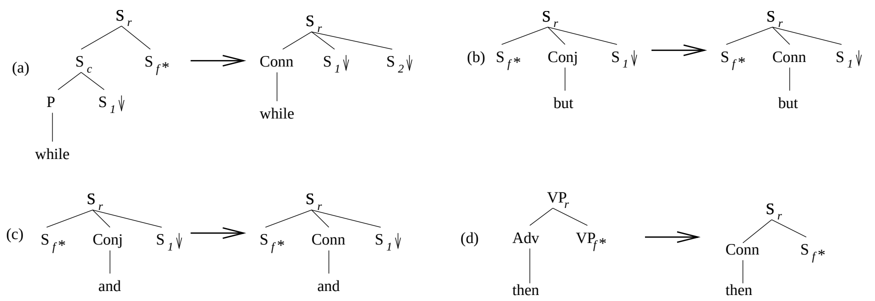

Figure 8. Elementary Tree Mappings for Connectives in Discourse (4)

DISCOURSE INPUT GENERATION. In the next phase of the system, the clausal derivations are first converted into elementary tree representations, which are treated as singular atomic units that can serve as the arguments of the discourse connectives. These clausal units, the input discourse, and the connective elementary trees generated by the Tree Mapper are then used to generate a discourse input representation that is essentially a sequence of lexicalized trees, where the lexicalizing elements are the connectives and the clausal units. Because of the extraction of the discourse units from the sentence derivations, and the tree mapping of the structures of connectives, tree selection ambiguities at the discourse level are minimized, and discourse parsing thus simplified.

This phase also includes an insertion algorithm to insert trees with an empty lexical anchor (which may still carry some feature information) into the input representation. Recall from Section 2 that the grammar contains an auxiliary tree that is used to continue the description by adjoining to the previous discourse (henceforth, continuation auxiliary trees) (Figure 3a). This auxiliary tree may be anchored by connectives like *and* and *or*, or remain lexically unrealized. In the extracted units shown in (5), there are only 2 overt connectives that can anchor this auxiliary tree: *and* and *but*. This means that the lexically empty trees need to be inserted at the appropriate positions in the input representation.We use the label "E" to indicate a null anchor. Alternatively, these trees can be taken to be lexicalized by the sentence-final punctuation markers.

The sequence of lexicalized trees after insertion of the lexically empty trees is then parsed with the same parser (LEM) that we used to parse each sentence in the discourse. Since the trees are uniquely selected by the connectives and the clausal units, the resulting discourse parse contains no ambiguities caused by tree selection.[10] However, the system does contain attachment ambiguities due to the continuation auxiliary trees. In the current approach, these may be resolved with an inferential component, or by statistical methods. For present purposes, we pick a unique derivation out of all the parses which satisfies the following two criteria: (a) adjunction in initial trees is only allowed at the root node; and (b) for all other permissible adjunctions, only lowest adjunction is allowed. Given the simple grammar posited in the system, these two criteria are sufficient to yield a unique derivation.

The derived tree and derivation structure for the example discourse in (4) after discourse parsing are shown in Figure 9. The current work does not discuss how the sematics for the discourse structure is derived after discourse parsing, but we hope to model the derivation of discourse semantics in a fashion similar to compostional semantics at the sentence level. See also Section 5.1.

We have also tested our system on connective rich sections of the Wall Street Journal (WSJ) from the Penn Treebank [19]. In order to avoid the problem of getting too many sentential derivations for the long and complex sentences typically found in this corpus, we used the single derivations produced by LexTract [38], which takes the Treebank and Treebank-specific information and produces derivation trees for the sentences annotated in the Treebank. For the WSJ discourse segment (taken from Section 21 of the WSJ corpus) given in Example 6,

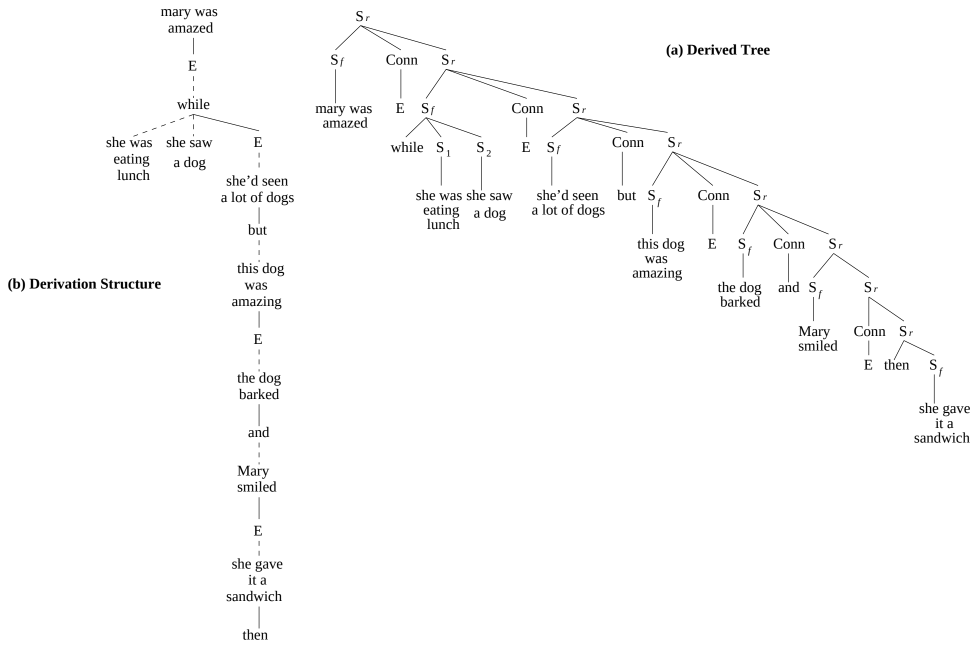

Figure 9. Derived Tree and Derivation Structure for Example Discourse in (4)

the derived tree and derivation structure are shown in Figure 10. The discourse connectives in the text are shown in bold.

(6) a. The pilots could play hardball by noting they are crucial to any sale or restructuring because they can refuse to fly the airplanes.[11]

b. **If** they were to insist on a low bid of, say $200 a share, the board mighn't be able to obtain a higher offer from other bidders **because** banks might hesitate to finance a transaction the pilots oppose.

c. **Also**, **because** UAL chairman Stephen Wolf and other UAL executives have joined the pilots' bid, the board might be able to exclude him from its deliberations in order to be fair to other bidders.

5. Discussion

5.1. Deriving Discourse Semantics

In [9] and [11], an approach to compositional semantics was provided for the LTAG grammar. The compositional semantics was defined with respect to the derivation tree structure and not the derived tree. The derivation tree is a record of the history of composition of the elementary trees. Semantic representations are associated with the elementary

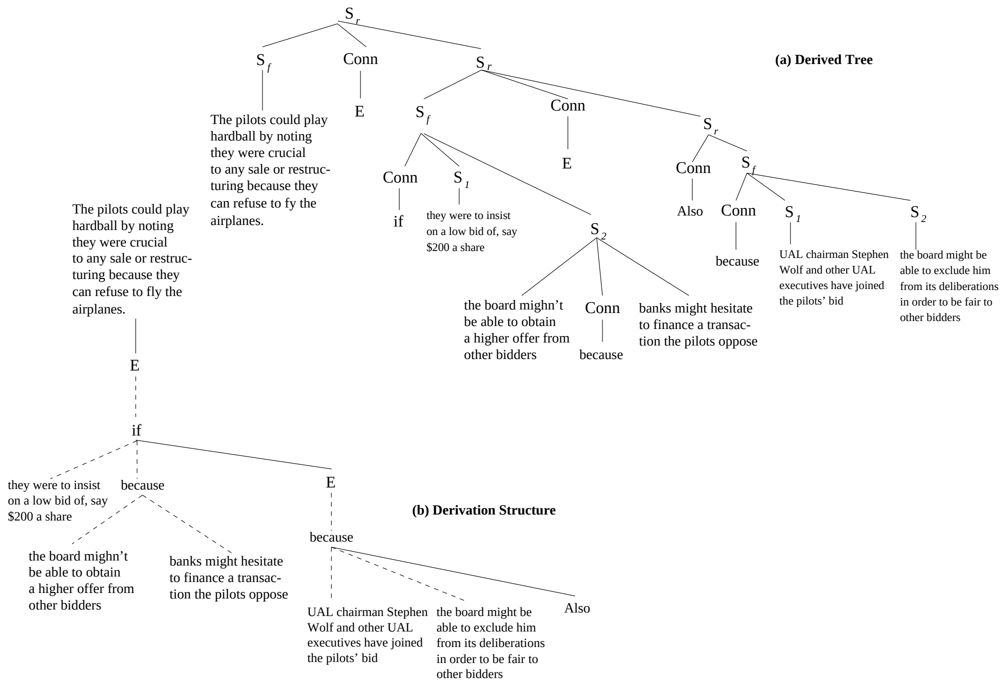

Figure 10. Derived Tree and Derivation Structure for WSJ Discourse in (6)

trees and these are composed via unification. In D-LTAG, the intuition is that a similar process will be carried out at the discourse structure level using the derivation tree of the D-LTAG grammar. The details of this compositional semantics for D-LTAG have not been worked out yet. However, in general, the final representation will be essentially a *flat* structure, as is the case for the semantics of LTAG.

While each elementary LTAG tree is associated with a semantic representation, this representation does not have to reproduce the hierarchical structure of the elementary tree. The elementary tree is thus considered as a "semantic unit". This view of representing semantics directly corresponds to the so-called flat representation, which is a conjunction of formulas, where hierarchical structure appears only when needed, for example, for operators on predicates. Such a "flat" representation is also motivated in the context of generation, where one wants to start with a representation of the input which makes the minimal commitment to structure. Details on associating a flat semantics with a derived TAG tree can also be found in [30]. In (7) below, we roughly illustrate the semantic formula associated with the example discourse in (4). We refer to the semantics of the connective trees by the names of the connectives, and use "&" to represent the semantics associated with the auxiliary tree associated with continuation/elaboration. The numbered arguments of these trees are labeled either "S" for states or "E" for events, depending on their semantic content.

(7) S1 & while(S2,E3) & but(S4,S5) & (E6 & E7) & after(E7, E8)

5.2. Discourse Connectives, Information Structure, and Discourse Semantics

While the current study does not directly address interactions between information structure (IS) and discourse structure/semantics, we note that a *lexicalised grammar* for both sentences and discourse allows semantic material from both the lexicon and constructed phrases to project into both sentence-level meaning and discourse meaning. In particular, this allows sentence-level IS distinctions to be projected into discourse-level descriptions. We illustrate this by considering clause-medial adverbial discourse connectives.

Many adverbial connectives display a wide variability with respect to the position they are found in the sentence. This variability, while usually not altering the compositional semantics of the sentence, appears to correlate with IS. In particular, we are pursuing the hypothesis that a clause-medial connective indicates that material to its left serves as a *contrastive theme*. The simplest case is given in Examples 8 and 9.

(8) Mary smiled. *However*, John frowned.

(9) Mary smiled. John, *however*, frowned.

In (8), the clause-initial position of *however* is, by itself, neutral about the partition of the sentence into *theme* or *rheme* and about whether or not the theme is contrastive. In (9), on the other hand, the clause-medial position of *however* correlates with stress on "John" and appears to convey that John and Mary are elements of an alternative set (in the sense of [22]) - that is, that John serves as a contrastive theme.

This comes out more strongly in Examples (10) and (11). (10) is infelicitous because medial *however* flags the subject as contrastive theme, but this subject is a coreferential (unstressed) pronoun and cannot serve as a contrastive theme. Example (11), on the other hand, is fine, as the position of *however* flags the adverbial *then* as the contrastive theme (presumably in an alternative set with the time of Mary smiling).

(10) * Mary smiled. Then she, *however*, frowned.

(11) Mary smiled. Then, *however*, she frowned.

Our claim here is just that, by having elements lexicalised both with respect to sentence and discourse, we can represent in the same way their contributions to both, as well as inter-relations between them. For example, in (11), not only does the clause-medial position of *however*, flag *then* as a contrastive theme (in contrast with alternatives provided in the discourse or the speech situation), but the defeasible rule presupposed (or conventionally implicated) by *however* ([13], [15]) involves that specific "inertial" property - i.e., if someone smiles, they will continue to do so. *However* asserts that it fails to hold, and what happened *then* is the source of the failure. While we have not yet explored this with respect to LTAG and D-LTAG, [1] and [2] illustrate how another lexicalised grammar, Combinatory Categorial Grammar (CCG) [27], can be used to express both assertional and presuppositional components of meaning associated with the sentence and with discourse, and [28] shows how one can compute both IS-partitioning, its prosody and its semantics in lockstep with other aspects of meaning.

6. Comparison with Related Approaches

Recently, [18] developed a system for identifying rhetorical relations on unrestricted text. His system trains on a corpus annotated with rhetorical relations and utilizes correlations of surface-based features with rhetorical structure relations to assign rhetorical structure to unseen text. Our system is a clear departure from this approach in two significant ways: a) we develop a system that actually parses discourse allowing the semantics to be built compositionally from the sentence to the discourse level, and b) discourse connectives are not viewed as names of relations, instead the semantics of the connectives form only a part of the compositional derivation of discourse relations. [12]

[4] uses a variant of Feature-based Tree Adjoining Grammars to constuct the structure of discourse and the semantics derived from it. [26] extends Gardent's formalism to handle world and contextual knowledge, proposing a non-monotonic reasoning system to achieve that. Despite this similarity of the above works with our approach, both systems differ significantly from ours in the following way. Gardent's system (also Schilder's) builds the semantics of discourse compositionally but only after the semantics of the input segments and the rhetorical relation connecting every two segments is identified. However, it is not clear how the semantics of the input segments are computed since, apparently, the size of the input segment ranges from tensed clauses ('We were going to take John as a lawyer'), to complex sentences ('As we found out, either he is on sick leave') or even fragments ('Too honest

for his own good, in fact').[13] In our approach, we do not assume preprocessing or segmentation of the textual input. The output from the sentence level parser is the input to the discourse parser, building up the semantics compositionally from the sentence level to the discourse level. Likewise, rhetorical relations are not assumed nor picked out from a previously defined set of relations. We are interested in those aspects of discourse interpretation that are *derived* compositionally, factoring away non-compositional semantic contributions, i.e. inferencing based on world-knowledge and anaphoric presuppositions.

7. Conclusions

Building on earlier work, we have developed and implemented a system for discourse parsing based on a lexicalized Tree-adjoining Grammar for discourse, in which the discourse connectives are the predicates, and the clauses are the arguments of these connectives. The system takes a discourse as its input, parses the sentences independently, extracts "discourse" connectives and clausal units from the output derivations of the sentences, and reparses the discourse input by submitting fully lexicalized trees to the same parser.

We have motivated a corpus study of discourse connectives in order to fully determine the semantic contribution they make to discourse, and thus, to also determine the elementary tree type(s) they lexicalize in the discourse grammar. The grammar thus developed will serve as a crucial component of the implemented system which uses this information after extracting the connectives from the sentence derivations, in order to create lexicalized elementary trees at the discourse-level. The submission of the lexicalized trees as the input for discourse level parsing simplifies the parsing process considerably, and this simplification is achieved because the system integrates sentence-level processing with discourse-level processing.

Acknowledgements

We thank Miriam Eckert, Alexandra Kinyon, Alistair Knott, Bangalore Srinivas, Fei Xia, and four anonymous reviewers (for the ESSLLI2001 workshop and for this volume) for their comments, which have helped in improving the content and presentation of this paper.

Notes

[1] Our use of the term *anaphora* does not include anaphoric relations such as those established by pronouns and definite descriptions. Obviously, a full account of the phenomenon of anaphora in discourse will have to take these into account. But they are not our present concern.

[2] In all the elementary trees shown in the paper, "$\diamond$" marks the anchor of the tree, "$\downarrow$" marks the substitution nodes, and "$*$" marks the adjunction nodes. Subscripts are used to distinguish non-terminal nodes with the same label.

[3] As noted in [33], both members of the pair need not be realized in the surface string.

[4] However, until we have accumulated ample empirical evidence, such conclusions are tentative and subject to revision. Also, it would be very interesting to investigate languages which allow crossed dependencies at the sentence level (e.g. Dutch) and examine whether in those languages crossed dependencies are also permissible on the discourse level. Our conjecture is that this will not be the case. The distinction between anaphoric and structural relations is discussed in greater length in [37].

[5] As an anonymous reviewer pointed out, it might be possible to account for such complex examples if we take into consideration the distinction between *semantic* and *pragmatic* relations holding between text spans. The literature on this distinction is summarized well in [14]. *Semantic* relations hold between the content of the text spans linked by the relation, whereas *pragmatic* relations hold between the utterances comprising the text spans, or the speech acts (or beliefs) underlying them. Of particular interest to our case is the observation that the same connective can be used for signalling either a semantic or a pragmatic relation.

[6] In this paper, we assume that clausal units correspond to the minimal tensed clause. The tensed clause is further taken to include all sentential complements, relative clauses and participial clauses. In some other discourse works, such as [20], a greater range of propositional elements are regarded as the minimal units of discourse.

[7] Each extracted clause derivation is taken to be an *atomic* unit in the discourse grammar, much like a single lexical item.

[8] In derivation structures, dotted lines indicate substitution and solid lines indicate adjunction. Also, note that each node is labeled by the lexical items, but these only serve as labels for the elementary tree with which they are associated.

[9] The hypothesis we are pursuing is that a clause-medial connective flags material to its left as being a contrastive theme [28] - cf. Section 5.

[10] This result obtains because the discourse grammar assumed here is quite simple, with discourse connectives projecting a single elementary tree. We note that upon further empirical investigation of the behavior of individual connectives, this may not turn out to be the case.

[11] Note that "because" in this sentence has not been treated as a connective because, initially, we have ignored sententially embedded connectives. How they contribute to discourse structure and meaning remains an important topic for future work. See also fn 7.

[12] In other words, in our view, the 'name' of a rhetorical relation is ultimately derived from the compositional semantics of our system, and other non-compositional aspects of discourse meaning, i.e. the inferential component. The use of 'rhetorical relations' in discourse intepretation seems to conflate those two distinct aspects of

meaning, namely compositional and inferential. In our system, we tease the two apart and derive the compositional part.

[13] The examples are from [4].

References

1. Bierner, G., and B. Webber: 2000. Inference through Alternative Set Semantics. *Journal of Language and Computation*, **Vol. 1, No. 2**, pp. 259-274.
2. Bierner, G.: 2001. Alternative Phrases and Natural Language Information Retrieval. *ACL Proceedings*, pp. 58-65.
3. Forbes, K. and E. Miltsakaki: 2002. Empirical Studies of Centering Shifts and Cue Phrases as Embedded Segment Boundary Markers. *Penn Working Papers in Linguistics, 'Current Work in Linguistics'* (ed.) Kaiser, E., **Vol. 7.2**, pp. 39-57.
4. Gardent, C.: 1997. Discourse Tree Adjoining grammars. *Univeristät des Saarlandes, Saarbrücken, Claus Report*, **No. 89**.
5. Grosz, B.J. and C. L. Sidner: 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, **Vol. 12, No. 3**, pp. 175-204.
6. Hirschberg, J. and D. J. Litman: 1987. Now Let's Talk About Now: Identifying Cue Phrases Intonationally. *ACL Proceedings*, pp. 163-171.
7. Hitzeman, J., M. Moens, and C. Grover: 1995. Algorithms for Analysing the Temporal Structure of Discourse. *EACL Proceedings*, pp. 253-260.
8. Hovy, E.H.: 1993. Automated Discourse Generation using Discourse Structure Relations. *Artificial Intelligence*, **Vol. 63, No. 1-2**, pp. 341-385.
9. Joshi, A. and K. Vijay-Shanker: 1999. Compositional Semantics with Lexicalized Tree-Adjoining Grammar (LTAG): How Much Underspecification is Necessary?. *IWCS Proceedings* (eds.) Blunt, H.C. and E.G.C. Thijsse, pp. 131-145.
10. Joshi, A. and S. Kuhn: 1979. Centered Logic: The Role of Entity Centered Sentence Representation in Natural Language Inferencing. *IJCAI-79*, pp. 435-439.
11. Kallmeyer, L. and A. Joshi: 1999. Factoring Predicate Argument and Scope Semantics: Underspecified Semantics with LTAG. *Proceedings of the 12th Amsterdam Colloquium*, (ed.) Dekker, P., pp. 169-174.
12. Kehler, A.: 2000. Resolving Temporal Relations using Tense Meaning and Discourse Interpretation. *Formalizing the Dynamics of Information*, (eds.) Faller, M., S. Kaufmann, and M. Pauly. CSLI publications.
13. Knott, A.: 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations.* PhD Thesis. University of Edinburgh, Edinburgh.
14. Knott, A.: 2001. Semantic and Pragmatic Relations and their Intended Effects. *Text Representation: Linguistic and Psycholinguistic Aspects*, (eds.) Sanders, T., J. Schilperoord and W. Spooren. Benjamins, pp. 127-151.
15. Lagerwerf, L.: 1998. *Causal Connectives have Presuppositions.* Holland Academic Graphics. The Hague, The Netherlands.
16. Lascarides, A. and A. Nicholas: 1993. Temporal Interpretation, Discourse Relations and Commonsense Entailment. *Linguistics and Philosophy*, **Vol. 16(5)**, pp. 437-493.
17. Mann, W.C. and S.A. Thompson: 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, **Vol., 8, No. 3**, pp. 243-281.

18. Marcu, D.: 2000. The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach. *Computational Linguistics*, **Vol. 26, No. 3**, pp. 395-448.
19. Marcus, M.P., B. Santorini, and M.A. Marcinkiewicz: 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, **Vol. 19, No. 2**, pp. 313-330.
20. Polanyi, L.: 1996. *The Linguistic Structure of Discourse.* CSLI Technical Report, **CSLI-96-200**.
21. Polanyi, L., and M.H. van den Berg: 1996. Discourse Structure and Discourse Interpretation. *Proceedings of the 10th Amsterdam Colloquium*, (eds.) Dekker, P., and M. Stokhof), pp. 113-131.
22. Rooth, M.: 1992. A Theory of Focus Interpretation. *Natural Language Semantics*, **Vol. 1**, pp. 75-116.
23. Sarkar, A.: 2000. Practical Experiments in Parsing Using Tree Adjoining Grammars. *Proceedings of the 5th TAG+ Workshop*, pp. 193-198.
24. Sarkar, A.: 2001,. Applying Cotraining Methods to Statistical Parsing. *Proceedings of the 2nd NAACL.*
25. Scha, R., and L. Polanyi: 1988. An Augmented Context Free Grammar for Discourse. *Proceedings of COLING*, pp. 573-577.
26. Schilder, F.: 1997. Tree Discourse Grammar, or How to Get Attached to a Discourse. *Proceedings of the Tilburg Conference on Formal Semantics.*
27. Steedman, M.: 2000. *The Syntactic Process.* MIT Press, Cambridge MA.
28. Steedman, M.: 2000. Information Structure and the Syntax-Phonology Interface. *Linguistic Inquiry*, **Vol. 34**, pp. 649-689.
29. Stone, M., and B. Webber: 1998. Textual Economy through Close Coupling of Syntax and Semantics. *Proceedings of the 9th INLG Workshop*, pp. 178-187.
30. Stone, M., C. Doran, and B. Webber: To appear. Microplanning from Communicative Intentions: Sentence Planning Using Descriptions (SPUD). *Computational Intelligence*, 2003.
31. Walker, M.A.: 1993. *Informational Redundancy and Resource Bounds in Dialogue.* PhD Thesis. University of Pennsylvania, Philadelphia.
32. Webber, B.: 1988. Tense as Discourse Anaphor. *Computational Linguistics*, **Vol. 14, No. 2**, pp. 61-73.
33. Webber, B., and A. Joshi: 1998. Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse. *Discourse Relations and Discourse Markers: Proceedings of the ACL/COLING workshop* (eds.) Stede, M., L. Wanner, and E. Hovy, pp. 86-92.
34. Webber, B., A. Knott, M. Stone, and A. Joshi: 1999. Discourse Relations: A Structural and Presuppositional Account using Lexicalised TAG. *ACL Proceedings*, pp. 41-48.
35. Webber, B., A. Knott, M. Stone, and A. Joshi: 1999. What are Little Trees Made of: A Structural and Presuppositional Account using Lexicalized TAG. *ACL Proceedings*, pp. 151-156.
36. Webber, B., A. Knott, and A. Joshi: 2000. Multiple Discourse Connectives in a Lexicalized Grammar for Discourse. *Collected papers form the 3rd IWCS.*
37. Webber, B., M. Stone, A. Joshi, and A. Knott: To appear. Anaphora and Discourse Structure. *Computational Linguistics*, 2003.
38. Xia, F., M. Palmer, and A. Joshi: 2000. A Uniform Method of Grammar Extraction and its Applications. *Proc. of the Joint Sigdat Conference on EMNLP/VLC.*
39. XTAG-Group, The: 2001. *A Lexicalized Tree Adjoining Grammar for English.* University of Pennsylvania, Technical Report, **IRCS-01-03**.