ELSEVIER

Cognitive Science 28 (2004) 751–779

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

D-LTAG: extending lexicalized TAG to discourse

Bonnie Webber

School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK

Received 27 May 2003; received in revised form 20 October 2003; accepted 26 April 2004

Abstract

This paper surveys work on applying the insights of lexicalized grammars to low-level discourse, to show the value of positing an autonomous grammar for low-level discourse in which words (or idiomatic phrases) are associated with discourse-level predicate–argument structures or modification structures that convey their syntactic-semantic meaning and scope. It starts by describing a lexicalized Tree Adjoining Grammar for discourse (D-LTAG). It then reviews an initial experiment in parsing text automatically, using both a lexicalized TAG and D-LTAG, and then touches upon issues involved in how lexico-syntactic elements contribute to discourse semantics. The paper concludes with a brief description of the Penn Discourse TreeBank, a resource being developed for the study of discourse structure and semantics.

Keywords: Discourse structure; Discourse connectives; Discourse parsing; Lexicalized grammar; Anaphora

1. Introduction

For many linguists, *syntax*—the structural regularities of a language that project the meanings of words onto those of utterances—stops at the sentence boundary. Material outside that boundary—i.e., the previous *discourse*—is simply context that may (or may not) license a particular construction of linguistic interest.

Of course, discourse too has structural regularities. Even in just the areas of formal and computational linguistics, there have been several attempts to produce a rigorous characterization of the regularities of discourse structure. For example,

- McKeown (1985) took the regularities she observed in the structure of definitions and encoded them into *schemata*, which could then be used to automatically generate definitions of concepts underlying a database model.

E-mail address: bonnie@inf.ed.ac.uk (B. Webber).

0364-0213/$ – see front matter © 2004 Cognitive Science Society, Inc. All rights reserved.
doi:10.1016/j.cogsci.2004.04.002

- Mann and Thompson (1988) observing regularities in the semantic and pragmatic relationships holding between adjacent clauses and taking them to hold recursively between larger units of discourse as well (i.e., clauses linked together by such relations), codified their observations about the resulting structures in a system called *Rhetorical Structure Theory* (RST). RST has provided an underpinning for work in Natural Language Generation (Hovy, 1988; Moore, 1990) and more recently, in summarization (Marcu, 2000).
- Grosz and Sidner (1986) focused on speaker *intentions* as a structuring principle for discourse, with a structural *dominance* relation holding between one discourse segment and those segments that supported its purpose, and a structural *precedence* relation between a discourse segment and those whose purposes required prior satisfaction.
- Sibun (1992) stressed the aleatory structure of expository discourse, as demonstrated in the descriptions of house and apartment layouts collected by Linde (1974), as well as similar descriptions she collected herself. Sibun showed how this structure could be modelled as the output of a semi-deterministic process reacting sequentially to properties of the world (viewed as potentially complex graph) that it was called upon to describe.
- Asher and Lascarides (1998) in *Segmented Discourse Representation Theory* focused on *reasoning* as an underpinning to discourse structure, explaining both discourse structure and the interpretation of discourse phenomena (e.g., anaphor resolution and presupposition grounding) as a by-product of reasoning (either monotonic or defeasible) about the way that a proposition connects to an accessible *speech act discourse referent* in the discourse context. Constraints on what is accessible mean that the resulting discourse has a tree structure.

While all these notions of structure apply to discourse, a more basic question—tied to syntax at the sentence-level—is whether such syntax *does* stop at the sentence boundary or whether the kind of syntactic regularities one sees at the phrase and sentence-level, that act with words to convey meaning, extend beyond the sentence into discourse.

Here we see work by Gardent (1997), Polanyi (1996) and Schilder (1997). They, like Asher and Lascarides (1998), were concerned with both discourse processing and discourse semantics—how each new segment of a discourse would be correctly attached to an evolving, interpreted discourse structure, such that the interpretation of the current structure was always available. Of particular interest here is that these researchers used the *adjoining* operation from Tree Adjoining Grammar (Joshi, 1987) and a related *sister-adjoining* operation in their work, as a way of constructing discourse structures *incrementally* from a sequence of sentences and clauses.

But these researchers did not explicitly address the way in which syntax might extend beyond the sentence, which is essentially the concern of the work that Aravind Joshi and I and some of our colleagues and students have been carrying out, in looking at *lexicalized grammars* for discourse.

In a lexicalized grammar, structure has a more intimate association with words than it does in a *phrase structure grammar*. For example, Lexicalized Tree Adjoining Grammar (Schabes, 1990) differs from a basic TAG in associating each entry in the lexicon with the set of tree structures that specify its local syntactic configurations.[1] Some of these tree structures can

combine with one another via *substitution*, while others make use of TAG's *adjoining* operation in order to produce a complete analysis (cf. Section 2).

In 1997, working with Dan Cristea (Cristea & Webber, 1997), I noticed that if one wanted to "parse" discourse incrementally in a TAG framework (following (Gardent, 1997; Schilder, 1997)), one also needed to exploit *substitution*, as well as the *adjoining* operation that they were already using. This was because it was necessary to associate a discourse connective such as "on the one hand" with a tree structure into which a subsequent, not necessarily adjacent, sentence marked by "on the other (hand)" or other contrastive marker, would then substitute, rather than adjoin. This brought the framework closer to a lexicalized TAG, and led Aravind Joshi and myself to begin to explore whether the insights of lexicalized grammars could also be applied to low-level discourse, that is, whether one could have an autonomous grammar for low-level discourse in which words (or in some cases, idiomatic phrases) were associated with discourse-level predicate–argument structures or modification structures that conveyed their syntactic-semantic meaning and scope (Webber & Joshi, 1998).

This exploration has continued over the last 6 years, engaging the attention and efforts of several students and colleagues (Creswell et al., 2002; Forbes, 2003; Forbes et al., 2001; Forbes & Webber, 2002; Forbes-Riley, Webber, & Joshi, submitted for publication; Miltsakaki, Creswell, Forbes, Joshi, & Webber, 2003; Miltsakaki, Prasad, Joshi, & Webber, 2004; Prasad, Miltsakaki, Joshi, & Webber, 2004; Webber, Joshi, & Knott, 2000; Webber, Knott, & Joshi, 2001; Webber, Knott, Stone, & Joshi, 1999a,b; Webber, Stone, Joshi, & Knott, 2003). Some of what we believe has been gained through this exploration is specific to a lexicalized approach to discourse, while other gains have been truly new and general insights into the way in which discourse structure and semantics project from lexico-syntactic elements. We hope the reader will grasp both sorts of benefits from this brief survey paper and from the papers it draws on. In particular, we hope to show that:

- The approach provides a uniform way for lexico-syntactic elements to contribute to the syntax and semantics of both the clause and discourse, opening up the (still to be realized) possibility of sentence processing and low-level discourse processing being carried out in an integrated fashion.
- The approach shows that low-level discourse structure and semantics is not simply a matter of attaching each new clause or sentence into the previous discourse through its discourse connectives: there are other ways in which discourse connectives can contribute to discourse coherence. These contributions can then interact, allowing certain complex features of discourse to be computed through the interaction of simpler mechanisms that are operational elsewhere as well.
- The approach allows one to reliably annotate a large corpus with low-level discourse structure, in which the basis for annotation decisions – discourse connectives (viewed as predicates) and their arguments – is clear.

The paper aims to demonstrate these benefits, surveying work carried out in this lexicalized approach to discourse and floating some new ideas as well. Section 2 illustrates what it means to have a lexicalized TAG for discourse – a D-LTAG – and how it relates to lexicalized TAG at the clause-level. It thereby shows how D-LTAG provides a uniform way for lexico-syntactic elements to contribute to both the clause and the discourse. Section 3 presents a brief look

at our first experiment on analysing discourse automatically with respect to D-LTAG (Forbes et al., 2001). This work uses the same chart-based left-corner LTAG parser (Sarkar, 2000) for both sentence and discourse processing, taking the sequence of derivation trees produced from sentence-level analysis and outputting a derivation tree for the discourse as a whole. It is a first step towards integrating sentence processing and low-level discourse processing.

Section 4 briefly describes how looking at text from a D-LTAG perspective – which requires one to associate a compositional semantic construction with each element of lexicalized syntax – has forced us to look more carefully at just how lexico-syntactic elements contribute to discourse semantics. The results are surprising: while some discourse connectives contribute a relationship between adjacent discourse elements as expected, others create an *anaphoric* relation between a discourse element and the discourse context. Still others, such as *for example* and *for instance*, contribute by abstracting over the nearest predication, be it clause-level or discourse-level, and adding the result to the discourse context (Webber et al., 2003). All of these – along with the ways they can interact – are described briefly in Section 4. Section 5 describes the *Penn Discourse TreeBank* (http://www.cis.upenn.edu/∼pdtb), a resource being developed for the study of discourse structure and semantics. Finally, Section 6 speculates on the future of D-LTAG.

It should be stressed that the focus of this work is properties of the *low-level* structure and semantics of monologic discourse. It does not address issues of *high-level* rhetorical structure (e.g., standard forms of narrative, argumentation or exposition), *intermediate-level* discourse structure in terms of speaker intentions, or *dialogue structure* (e.g., question–answer patterns, patterns of clarification dialogues or of exposition and acknowledgement, etc.). Thus, it does not pose an alternative to theories of intermediate- or high-level discourse structure or dialogue structure, but rather a necessary substrate for such theories, similar to that of sentence-level syntax and semantics.

2. D-LTAG: lexicalized TAG for discourse

In a lexicalized Tree-Adjoining Grammar (LTAG), a word is associated with a set of tree structures (its *tree set*), one for each *minimal syntactic construction* in which the word can appear. For example, within the tree set of *like* is one tree (Fig. 1a) corresponding to simple SVO order for transitive verbs, as in *The boys like apples*, another tree corresponding to topicalized OSV order (Fig. 1b), as in *Apples the boys like*, and a third tree corresponding to the simple passive (Fig. 1c), as in *Apples are liked by the boys*. All these trees realize the same predicate–argument structure, with one NP argument for the "liker" and a second NP argument for the "likee". The tree set also includes a tree corresponding to *like* as the prepositional head of an NP post-modifier (Fig. 1d), as in *apples like this one* and another tree corresponding to *like* as the prepositional head of a VP post-modifier (Fig. 1e), as in *Sing like a bird*.

The above syntactic/semantic encapsulation is possible because of the extended domain of locality of a lexicalized grammar: When *like* is simply characterized as a verb (or a preposition or a noun) in a non-lexicalized grammar, the information about the syntactic configurations it can appear in and how its interpretation combines with that of other elements in those syntactic

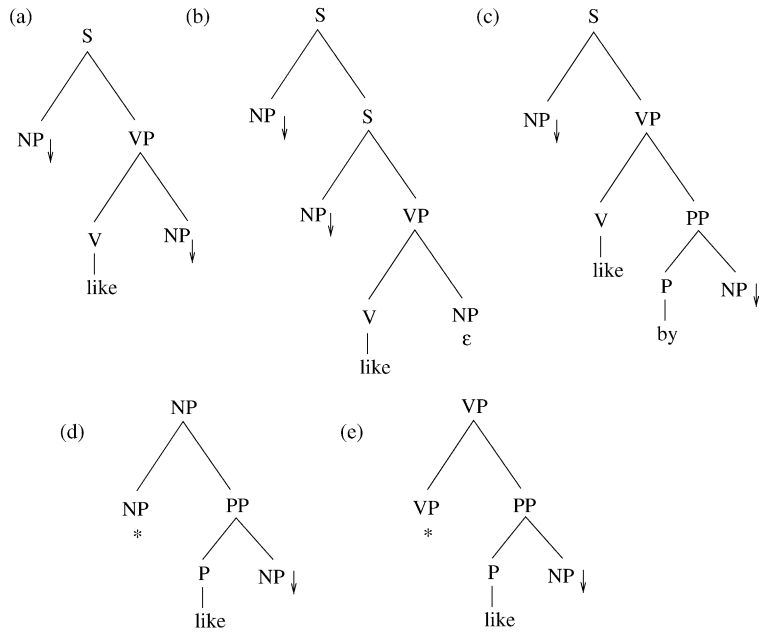


Fig. 1. Elements of the tree set of *like*.

configurations is spread out across other parts of the grammar rather than being localized in one place.

In an LTAG, there are two kinds of *elementary* tree structures that can appear in a tree set: *initial* trees that reflect basic functor-argument dependencies and *auxiliary* trees that introduce recursion and allow elementary trees to be modified and/or elaborated. Fig. 1a–c are all *initial* trees, while (d) and (e) are *auxiliary* trees. The special symbols used in these trees (↓ and *) relate to the two operations by which trees can combine to form more extended *derived* trees. ↓ indicates a *substitution site* where an elementary tree can substitute into a derived tree, provided the label at its root matches that of the substitution site. For example, an NP tree anchored by the proper noun *John* can substitute into any of the substitution sites in Fig. 1. * indicates an *adjunction site* (or *foot node*), where an auxiliary tree can adjoin into a root, leaf or non-terminal node of an elementary or derived tree, again provided that its label (the same as that of its root node) is also the same as the label of the node to which it is being adjoined. Fig. 2 shows the *like* PP tree from Fig. 1d and a tree corresponding to "John ate an apple", along with the result of adjoining the first tree into the second at its second NP node. Additional examples of adjoining can be found throughout the paper, as well as in other papers in this volume.

Now, one way of projecting the insights of lexicalized grammar into discourse would be to have a single grammar that mapped lexical items into discourse structures directly.[2] Such a radical step would not be impossible. However, we have not thought through its many consequences in detail, given that one would not want to lose the generalizations that have been captured over many years of work in lexicalized sentence-level grammars.

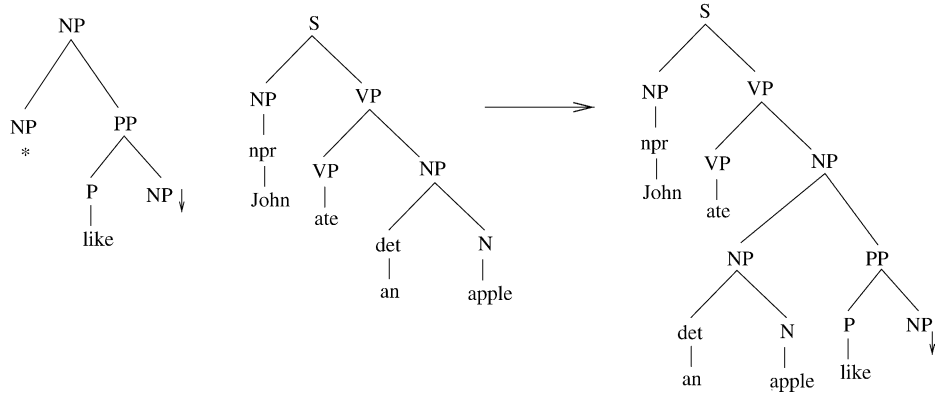


Fig. 2. An auxiliary PP tree adjoining to an initial NP tree.

Instead, we have simply posited a separate LTAG for discourse (D-LTAG) that uses the same operations of *substitution* and *adjoining*. While there is some overlap between the two (e.g., both providing an analysis of subordinate clause—main clause constructions, there one striking difference: While LTAG requires a wide variety of different elementary trees to describe clause-level structure, we have found that D-LTAG requires very few elementary tree structures, possibly because clause-level syntax exploits structural variation in ways that discourse doesn't. For example, *like* is associated in the XTAG grammar (XTAG-Group, 2001) with 28 elementary trees such as Fig. 1a–c in which it serves as a verb anchor. In contrast, a subordinate conjunction such as *because*, which in D-LTAG is a discourse-level predicate that takes two (clausal) arguments, is associated with only two elementary trees—the same two as every other subordinate conjunction. Thus, all the elementary trees so far identified as being needed for D-LTAG are presented in this short section.

The root node of an elementary tree in D-LTAG is a *discourse clause* (D_c). At each substitution site, a basic clause can be substituted or a derived tree. (A basic clause is treated as an atomic unit with features, just as word or lemma is in a sentence-level grammar.[3]) Other leaves are adjunction sites or the lexico-syntactic elements that anchor the tree. Here we will first look briefly at *initial* trees in D-LTAG and the range of lexical items that anchor them and thereby serve as the predicate of discourse-level predicate–argument structures (Section 2.1). We then look at *auxiliary* trees in D-LTAG and the lexical items that anchor trees that elaborate the ongoing discourse (Section 2.2).

2.1. Initial trees in D-LTAG

D-LTAG associates initial trees with a variety of lexico-syntactic elements that serve as predicates on clausal arguments: subordinate conjunctions and other *subordinators*; the lexico-syntactic anchors of parallel constructions; some coordinate conjuctions; and even some specific verb forms.

In LTAG (XTAG-Group, 2001), subordinate conjunctions such as *if*, *although*, *since* and *so that* anchor *auxiliary* trees because they are outside the domain of locality of the verb, heading

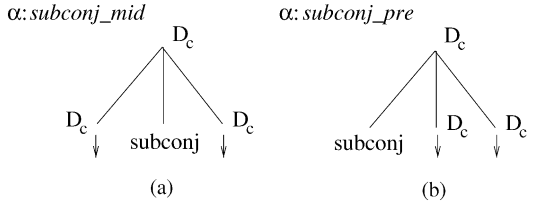


Fig. 3. Initial trees (a–b) for a subordinate conjunction.

clausal or VP *adjuncts*. In D-LTAG, however, it is predicates on clausal arguments that define the domain of locality. Thus, at the discourse-level, subordinate conjunctions anchor *initial trees* into which clauses substitute as arguments. Fig. 3 shows the initial trees for postposed subordinate clauses (a) and preposed subordinate clauses (b). In this and other figures, $\mathbf{D}_c$ stands for "discourse clause", ↓ indicates a substitution site, and <subconj > stands for the particular subordinate conjunction that anchors the tree.

Similar to subordinate conjunctions are what Quirk et al. (1972) call *subordinators*—lexical items such as *in order for*, *in order to*, and *to* (which head *purpose clauses*) and *by* (which heads a *manner clause*). These also anchor *initial* trees in D-LTAG, while anchoring *auxiliary* trees in LTAG. They only differ from subordinate conjunctions in having a non-finite (untensed) clause as one argument and a finite (tensed) clause as the other one.

D-LTAG also associates initial trees with the lexical anchors of parallel constructions such as

(1) *On the one hand*, John is generous. *On the other hand*, he's hard to find.

The initial tree for this parallel construction is shown in Fig. 4. It is associated with both the lexical anchors *on the one hand* and *on the other (hand)*. While in LTAG, these idiomatic prepositional phrases would anchor separate auxiliary trees that adjoin at the sentence-level, in D-LTAG, they both serve as anchors for the same initial tree, keeping the two discourse clauses (D_c) that substitute in, within the same domain of locality. There are similar multiply-anchored initial trees for *disjunction* ("either"…"or"…), *addition* ("not only"…"but also"…), and *concession* ("admittedly"…"but"…).

There are also initial trees anchored by coordinate conjunctions that convey a particular relation between the connected units, such as *so*, conveying *result*. Its initial tree is shown in Fig. 5. In contrast, we take the coordinate conjunction *and* to anchor an auxiliary tree, as discussed in the next section.

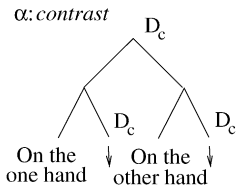


Fig. 4. Initial tree for a parallel contrastive construction.

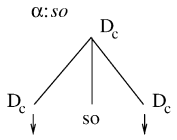


Fig. 5. Initial tree for the coordinate conjunction *so*.

Finally, there is motivation for taking the imperative form of *suppose* as anchoring an initial tree in D-LTAG. This differs from LTAG, where verbs such as *suppose* that take sentential complements are taken to anchor an *auxiliary* tree rooted in an S-node, as shown in Fig. 6a. This analysis provides a natural way for LTAG to handle the syntactic phenomenon of *long-distance extraction*(XTAG-Group, 2001), illustrated in sentences such as "Who does John suppose likes beans?" (where *who* is the subject of *likes*) and "Who did the elephant think the panda heard the emu say smells terrible?", where *who* is the subject of *smells*. With respect to this clausal analysis, the auxiliary tree for *suppose* in Example 2, would adjoin to the root of the tree for the clause "an investor wants to sell a stock . . .".

(2) Suppose an investor wants to sell a stock, but not for less than 55. A limit order to sell could be entered at that price.

At the discourse level, the motivation for taking imperative *s*uppose to anchor an *initial tree* with two substitution sites (Fig. 6b), is that it corresponds more closely to its discourse-level predicate–argument structure: One substitution site will be filled by its sentential complement, which specifies a hypothetical or counterfactual condition, while the second will be filled by a subsequent discourse clause which should be evaluated under that condition—here, "A limit order could be entered at that price". This is equivalent to the discourse-level predicate–argument structure of the subordinate conjunction *if*. As with *if*, the second argument of *suppose* need not be an assertion. It can instead be a command (Example 3) or a question (Example 4), as in these examples returned from Google:

(3) Suppose that the market is semi-strong form efficient, but not strong form efficient. Describe a trading strategy that would result in abnormally high expected returns.

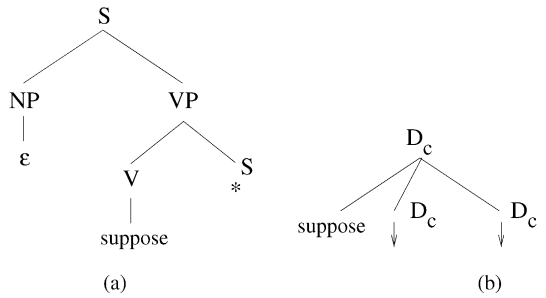


Fig. 6. LTAG and D-LTAG trees for imperative *suppose*.

(4) Suppose that you want to send an MP3 file to a friend, but your friend's ISP limits the amount of incoming mail to 1 MB and the MP3 file is 4 MB. Is there a way to handle this situation by using RFC 822 and MIME?

Of course, imperative *suppose* doesn't always play this discourse role, which leads to ambiguity in D-LTAG analyses as to whether a particular token of *suppose* projects an *initial* tree into the discourse, or just anchors a simple discourse clause, I will mention other sources of ambiguity in the next section.

One final point here. In all our D-LTAG papers to date, we have talked as if words anchor both LTAG trees and D-LTAG trees. Because it is often the case (as with *suppose*) that only when a lexical item occurs in a particular structural configuration that it should be associated with a particular tree in D-LTAG, it is more accurate to talk in terms of anchored LTAG trees anchoring D-LTAG trees. This is, in fact, how our initial parser for D-LTAG operates, as will be described in Section 3.

2.2. Auxiliary trees in D-LTAG

Auxiliary trees in an LTAG introduce recursion and allow elementary trees to be modified and/or elaborated. Auxiliary trees in D-LTAG do the same (Webber et al., 1999a,b, 2003). Here we describe the auxiliary trees that we have taken to be part of D-LTAG and then reflect on the justification for these decisions.

The first use of auxiliary trees in D-LTAG is in connection with descriptions of objects, events, situations and states that extend over several clauses in a discourse. Such extended descriptions are formed with coordinate conjunctions and/or unrealized (null) connectives. Thus, D-LTAG has taken both coordinate conjunctions and null connectives to anchor *auxiliary* trees—cf. Fig. 7a. When such a tree is adjoined to a discourse clause and its substitution site is filled with another discourse clause, the latter extends the description of the situation or entity conveyed by the former.[4] Such auxiliary trees are used in the derivation of simple discourses such as (5):

(5) a. John went to the zoo.
b. He took his cell phone with him.

This derivation is shown in Fig. 8. To the left of the arrow (→) are the elementary trees to be combined: T1 stands for the LTAG tree for clause 5a, T2 for clause 5b, and β:*unrealized*, for the auxiliary tree that connects adjacent clauses without an overt connective. In the derivation,

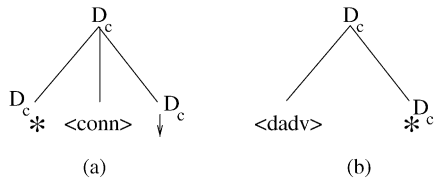


Fig. 7. Auxiliary trees in D-LTAG. <conn> stands for any explicit coordinating conjunction or null connective. <dadv> stands for any discourse adverbial.

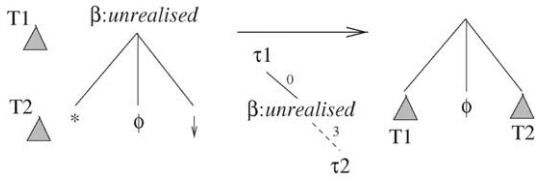


Fig. 8. D-LTAG derivation of Example 8.

the foot node of β:*unrealized* is adjoined to the root of T1 and its substitution site is filled by T2. The result is shown to the right of →. (A standard way of indicating TAG derivations is shown under →, in the form of a *derivation tree* in which solid lines indicate adjunction, and dashed lines, substitution. Each line is labelled with the address of the argument at which the operation occurs. τ1 is the derivation tree for T1, and τ2 is the derivation tree for T2.)

We have posited a second type of *auxiliary* tree for D-LTAG, shown in Fig. 7b. This one is anchored by a discourse adverbial such as *instead*, *otherwise*, *then*, *in contrast*, *therefore*, *for example*, *nevertheless*, etc. What is striking about this tree is that it is associated with only a *single* discourse clause, while both the initial trees in Figs. 3–6 and the auxiliary tree in Fig. 7a are associated with two distinct discourse clauses. That only a single discourse clause is involved in this second auxiliary tree (Fig. 7b) follows from our argument (Webber et al., 2003) that discourse adverbials, by and large, establish an *anaphoric* link between the interpretation of the clause to which they adjoin and the previous discourse.

But note that adverbials such as these could be interpreted with respect to the discourse *without* being distinct elements of discourse grammar, as is the case with the demonstrative pronouns ("this" and "that"): While a demonstrative pronoun is taken to refer to an *abstract object* evoked by the previous discourse (Asher, 1993; Webber, 1991), in subject or object position, it is part of the predicate–argument structure of the *verb*, so would not automatically be part of the discourse grammar. (In other positions, demonstrative pronouns contribute to adjuncts on the verb, but that does not make them automatically part of the discourse grammar either.)

Now in LTAG, the reason that adverbials anchor *auxiliary* trees is because they are outside the predicate–argument structure of the verb, contributing modifiers like *manner of action* (e.g., "swiftly"), *frequency* of actions or events (e.g., "annually"), *speaker attitude* towards events or situations (e.g., "unfortunately"), etc.[5] If one took the comparable position in D-LTAG, then discourse adverbials would anchor auxiliary trees in the discourse grammar if they were outside the predicate–argument structure of any nearby discourse predicate, i.e., any structural connective (including the null connective) or other discourse adverbial.

So there are two questions: (1) Should discourse adverbials, which are interpreted with respect to discourse in a way that clausal adverbials are not (Forbes, 2003), be treated as part of the discourse grammar and (2) if they should, is it auxiliary trees that they anchor?

Discourse adverbials like *instead* and *otherwise* belong in the discourse grammar because they related two *abstract objects* in the same way as do clausal adjuncts, as in Example 6:

(6) a. *Instead of staying home*, John went to the zoo.
b. *After cleaning the snow off his car*, John went to the zoo.
c. *Because he felt bored at home*, John went to the zoo.

The only difference in Example 7 below is that one of the *abstract object* arguments to *instead* is provided anaphorically—in this case, from the clausal subject of the previous sentence. Hence, we take discourse adverbials to belong to discourse grammar as well as to sentence-level grammar.

(7) Going to the beach sounded boring. *Instead*, John went to the zoo.

As for the second question, it is possible that these discourse adverbials should be taken to anchor an *initial tree* (as do subordinate conjunctions), but one whose first argument must be recovered anaphorically. Discourse adverbials like *for example* and *for instance* show why the projected structures should be taken to be auxiliary trees. In Webber et al. (2003), we show that these adverbials can operate on discourse predicates, as in Example 8:

(8) John broke his arm, so for example, he can't cycle to work now.

Here, the structural connective *so* is interpreted as relating the interpretation of "John broke his arm" and "he can't cycle to work now"—the latter being a consequent of the former. The discourse adverbial *for example* modifies the *extent* of the consequence—the latter being but one example of the consequences of the former. So these adverbials serve as adjuncts to discourse predicates, and hence as anchors for auxiliary trees. Notice, of course, that the predicate need not be explicit, as in Example 8:

(9) You shouldn't trust John. For example, he never returns anything.

Here, one infers that John's lying is meant to be an *explanation* for why one shouldn't trust him, with *for example* modifying its *extent*—that it's only one of possibly many reasons.

Nothing else in this paper depends on whether discourse adverbials should be modelled as auxiliary trees in both sentence-level LTAG and discourse-level D-LTAG, but the reader should be aware that it is a question whose answer tells upon how one thinks about discourse grammar.

I turn now to the topic of lexical ambiguity in D-LTAG, noting that there are other sources of lexical ambiguity beyond those mentioned in Section 2.1. One is associated with the fact that adverbials can appear in one structure in which they are discourse adverbials (depending on the discourse for part of their interpretation), as in 10a–b, and in other structures in which they are independent of the discourse, as in 10c–d.

(10)
a. *Instead*, John ate an apple.
b. *Otherwise*, you can forget dessert.
c. John ate an apple *instead* of a pear.
d. Mary was *otherwise* occupied.

In these cases, the clause-level analysis serves to disambiguate whether or not the lexical item functions at the discourse level.

Another source of ambiguity is invisible at the clause level. It stems from the fact that many of the adverbials found in second position in parallel constructions (e.g., *on the other hand*, *at the same time*, *nevertheless*, *but*) can also serve as simple discourse adverbials on their own. In the first case, they will be one of the two anchors of an initial tree, such as in Fig. 4, while in the second, they will anchor the simple auxiliary tree shown in Fig. 7b. This lexical ambiguity

leads to *local ambiguity* at the discourse level. That is, while there is only one consistent *global* analysis of the discourse, an incremental parser, working left-to-right, faces a choice that can only be decided based on material that comes later. This is something that clause-level parsers face on a regular basis.

For example, in the following passage, *at the same time* serves as the second anchor of an initial tree expressing contrast, whose first anchor is *on the one hand*.

(11) Brooklyn College students have an ambivalent attitude toward their school. *On the one hand*, there is a sense of not having moved beyond the ambiance of their high school. This is particularly acute for those who attended Midwood High School directly across the street from Brooklyn College. . . .*At the same time*, there is a good deal of self-congratulation at attending a good college . . ..

However, in the following minor variation of Example 11, *at the same time* anchors an auxiliary tree that elaborates on the positive aspects of attending Brooklyn College, with *on the other hand* serving as the second anchor of the initial tree that expresses contrast.

(12) Brooklyn College students have an ambivalent attitude toward their school. *On the one hand*, there is a good deal of self-congratulation at attending a good college. *At the same time*, they know they're saving money by living at home. *On the other hand*, there is a sense of not having moved beyond the ambiance of their high school.

D-LTAG analyses do not introduce any kind of local or global discourse ambiguity that is not present in the original discourse. As with ambiguity at the clause-level, discourse ambiguity is a problem that parsers must punt on or deal with, as I will discuss briefly in the next section. As with clause-level ambiguity, discourse ambiguity is a problem that will probably be best solved by parsers using a combination of statistics (favoring analyses with the highest priors and textual evidence) and discourse semantics (favoring analyses that make referential and relational sense in the current context). All such work is in the future.

3. A parser for D-LTAG

Discourse parsing involves analyzing a discourse according to a discourse grammar—in our case, D-LTAG. To date, we have carried out a single experiment with discourse parsing (Forbes et al., 2001) that shows that the same parser can be used for both clause-level LTAG and D-LTAG. While it does not pretend to have any psycholinguistic validity, it does bring up some aspects of discourse processing worth commenting on further.

In this work, a chart-based left-corner LTAG parser, LEM(Sarkar, 2000) makes two passes through the text, the first producing XTAG derivation trees for each sentence from the sequence of elementary trees associated with its words, the second producing a D-LTAG derivation for the discourse as a whole from the sequence of elementary trees associated with its discourse connectives and clausal derivations. The flow of processing is shown in Fig. 9.

For each sentence in the discourse, LEM uses its chart to record possible *derivation trees* for the sentence according to the XTAG grammar (XTAG-Group, 2001). To produce a single analysis, heuristics can be used to decide which elementary tree to assign to each word (to

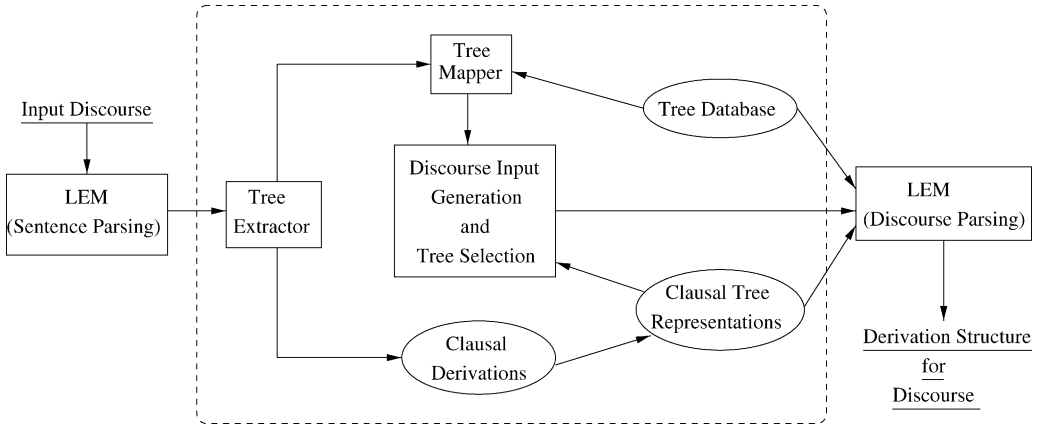


Fig. 9. Two-pass sentence/discourse parsing using LEM.

deal with *lexical* ambiguity), and to choose where to attach modifiers (currently, the lowest attachment point) to deal with *structural* ambiguity. Eventually, statistics will replace heuristics in this process.[6]

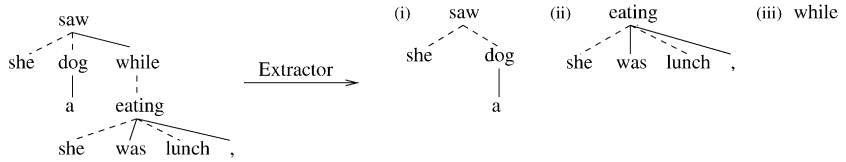
The sequence of derivation trees corresponding to the sequence of sentences in the discourse is input to a *Tree Extractor* (**TE**), which extracts two sorts of things from each one: (1) the derivation tree for each clause in the sentence, and (2) each elementary tree anchored in a discourse connective. This is done in two passes—the first, to identify the discourse connectives, and the second, to detach clausal derivations from their substitution and/or adjunction nodes. The first – a top-down traversal of the derivation tree – considers both *lexical* and *structural* properties of each lexical item because, as noted earlier,

- lexical items that can serve as discourse connectives can also be used in other ways (e.g., *instead* can serve as an NP post-modifier—"an apple instead of a pear"; *and* can serve as an NP conjunction). So lexical features alone are insufficient to determine whether a particular token is actually serving as a discourse connective in a particular context.
- LTAG does not distinguish between clausal adverbials like *frequently* and discourse adverbials like *otherwise*. So structural features alone are also insufficient.

So from the sentence

(13) While she was eating lunch, she saw a dog.

TE extracts the two clausal derivations and one elementary tree anchored in a discourse connective shown below.



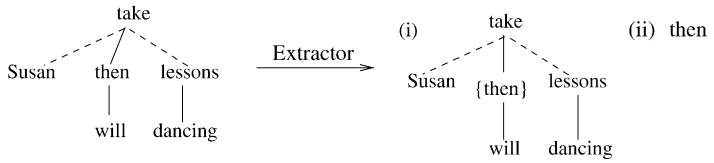


Fig. 10. Application of **TE** to the derivation tree of Example 14.

With clause-medial discourse connectives, as in

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

TE makes a *copy* of the derivation and replaces the discourse connective with an *index*, to retain its clause-internal position. This is because clause-medial adverbials appear to be relevant to Information Structure (Steedman, 2000), and thus their position in the clause is important to preserve.[7] So in Example 14, **TE** extracts a single clausal derivation and one elementary tree anchored in a discourse connective, as shown in Fig. 10.

Tree Mapping applies to the output of Tree Extraction, to map *sentence-level* structural descriptors of connective elementary trees to their *discourse-level* structural descriptors. (Note that this embodies the suggestion at the end of Section 2.1 that it is not lexical items that anchor D-LTAG trees, but rather anchored LTAG trees, e.g., only *otherwise* as an S-adjoining adverbial, and not as an adjective-adjoining adverbial.)

The role of the next stage of the process, Discourse Input Generation (**DIG**) is to produce a sequence of lexicalized trees, which can be submitted for publication to LEM for discourse parsing. The sequence of lexicalized trees consists of the connective elementary trees obtained from **Tree Mapping** and the clausal elementary trees corresponding to the clausal derivations obtained from the **Tree Extractor**. When there is no structural connective between clausal units, **DIG** inserts an auxiliary tree with an empty lexical anchor into the input sequence.

Ambiguity is handled at the discourse level much in the same way as at the clause level—a single tree is chosen for each connective and the lowest attachment point is selected. (In addition, adjunction in initial trees is only allowed at their root node.) Lowest attachment heuristics are illustrated in Example 15. The reason for selecting this example is that the interpretation of *they* in the final sentence seems to vary with the analysis selected, and so can be used as a diagnostic for that process.

(15) John is stubborn. (T1)
His sister is stubborn. (T2)
His parents are stubborn. (T3)
So they are continually arguing. (T4)

Fig. 11 shows the output from **DIG** for this example. The five possible derivations for this example are shown in Fig. 12, corresponding to five derived structures shown in Fig. 13.

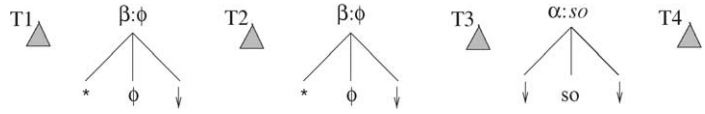


Fig. 11. Trees that serve as input to LEM's discourse parsing from Example 14. "John is stubborn. His sister is stubborn. His parents are stubborn. *So* they are continually arguing."

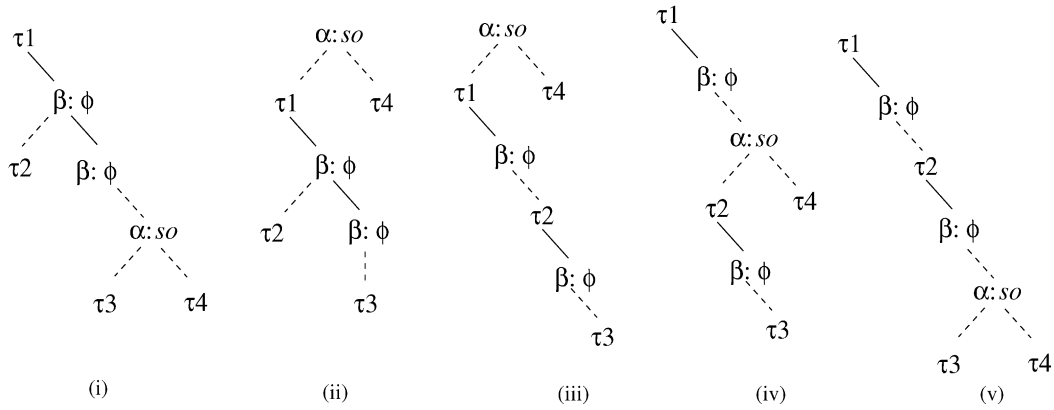


Fig. 12. Potential discourse-level derivation trees for Example 15.

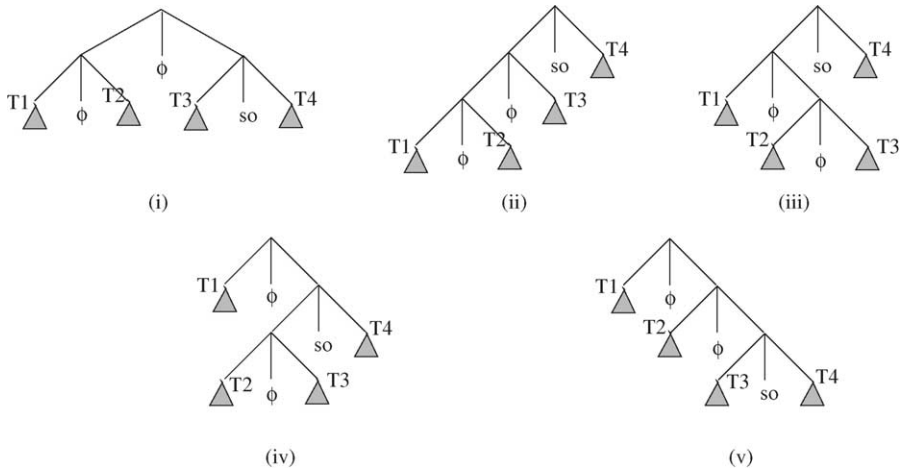


Fig. 13. Derived structures for discourse parsing of Example 15.

Structure (i) can be paraphrased as

> John and his sister are stubborn. His parents are stubborn. So they [his parents] are always arguing.

Structure (iv) can be paraphrased as

> John is stubborn. His sister and his parents are stubborn. So they [his sister and his parents] are always arguing.

while structures (ii), (iii) and (v) can all be paraphased as

> John and his sister and his parents are stubborn. So they [the whole family] are always arguing.

Most readers will take either this or the interpretation associated with structure (i) as the correct interpretation of Example 15, while having no feeling as to which of the structures has given rise to it. Our discourse parser, however, only considers the unique derivation in which (i) for an initial tree, adjunction is only allowed at the root node, while (ii) for all other trees, only the lowest adjunction is allowed. This means that the discourse parser only produces derivation (v) and derived tree (v) for Example 15, which happily accords with the one that most readers can get. Nevertheless, a more robust treatment of both lexical and structural ambiguity should be pursued.

There is one more problem that a parser for discourse must address—that of discourse embedded in indirect speech or a propositional attitude, as in (16) and (17).

(16) The pilots could play hardball by noting that they are crucial to any sale or restructuring because they can refuse to fly the airplanes.

(17) Epigenesists believed that the organism was not yet formed in the fertilized egg. Rather, it arose as a consequence of profound changes in shape and form during the course of embryogenesis.

In both examples, the sentential complement of the verb (*note* in (16) and *believe* in (17)) must itself be analysed as a discourse, extending in the case of (17) to the next sentence as well.

Our initial solution to this problem resembles, in part, our treatment of imperative *suppose* in Example 2. I have already mentioned, in discussing imperative *suppose*, that in LTAG, verbs that take sentential complements do so in the form of an *auxiliary* tree that adjoins to the object clause (cf. Fig. 6a). In D-LTAG however, we posit an *initial* tree for imperative *suppose* that takes two discourse clauses as arguments. For indirect speech and propositional attitude verbs, we are following a suggestion from Aravind Joshi and positing something similar: an *initial* tree anchored by the propositional attitude or indirect speech verb that has a covert argument that is coindexed with the (overt) clausal complement introduced by the complementizer (Fig. 14). So,

(18) John believes that Mary is tired.

is analyzed as

(19) John believes X_i that [Mary is tired]$_i$.

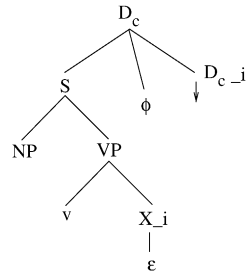


Fig. 14. Proposed D-LTAG initial tree for prepositional attitude and indirect speech verbs.

There is cross-linguistic evidence for such an analysis coming from Hindi,[8] where the X_i may be overtly expressed, as in

(20) raam ye samajhtaa hai ki sita thakii-huii hai
Ram this believes is that Sita tired is
Ram believes this that Sita is tired

The discourse analysis of Example 16 would then involve the trees shown in Fig. 15, where T1 represents the analysis of "The pilots could play hardball", T2 represents the analysis of "they are crucial to any sale or restructuring", and T3, the analysis of "they can refuse to fly the airplanes".

Similarly, the discourse analysis of Example 17 would involve the trees shown in Fig. 16, where T1 represents the analysis of "the organism was not yet formed in the fertilized egg" and T2, the analysis of "it arose as a consequence of profound changes . . .".

Neither this view of propositional attitude and indirect speech verbs, nor imperative *suppose*, nor the (local) ambiguity caused by discourse connectives that can appear in more than one D-LTAG tree, have yet been incorporated into the parser described earlier. I expect that when they are, we will discover other aspects of low-level discourse analysis that need exploring.

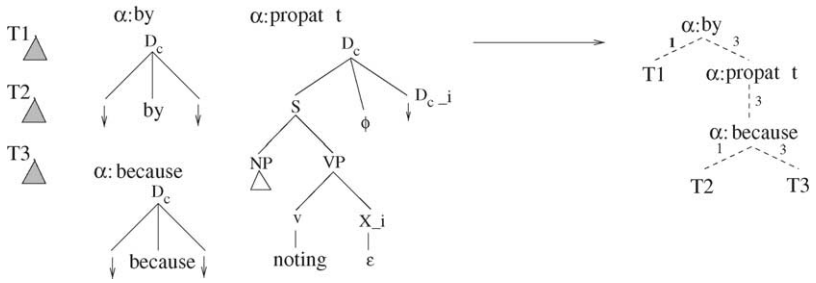


Fig. 15. D-LTAG derivation of Example 16 "The pilots could play hardball by noticing that they are crucial to any sale or restructuring because they can refuse to fly the airplanes."

4. Differences between discourse connectives in D-LTAG

As shown in Section 2, D-LTAG distinguishes between (1) structural connectives that anchor *initial* trees and convey discourse-level predicate–argument relations; (2) structural connectives (including the null connective) that anchor *auxiliary* trees and that elaborate the preceding discourse; and (3) discourse adverbials that anchor *auxiliary* trees and contribute predicate–argument relations distinct from (but that may interact with) those conveyed by structural connectives.

Webber et al. (2003) argue extensively that while structural connectives and discourse adverbials may both convey discourse-level predicate–argument relations, they get their arguments in different ways. Structural connectives get both their arguments from the discourse clauses to which they are structurally connected in the discourse, as in the following[9]

(21) a. **Because** [Healthcare actually owes HealthVest 4.2 million in rent and mortgage payments each month], [the amount due above the amount paid will be added to the three-year note.]

b. Even though critical, [it was just the kind of attention they were seeking.] **So** [they fired back at the Goldman Sachs objections in their own economics letter, "The BMC Report."]

On the other hand, many discourse adverbials get only one argument from the clause or sentence to which they are adjoined and the other anaphorically from the preceding discourse as in

(22) a. [If the light is red], stop. **Otherwise**, [continue down the road.]

b. One great difference distinguished the Soviet and German systems: [there was no Soviet equivalent of the death camps]. People sentenced to death in the Soviet Union were generally shot before entering the camp network. Applebaum estimates these victims at just under one million during the Stalin years. **Instead**, [Soviet prisoners were expected to earn their keep by contributing to the creation of Soviet Socialism]

c. A person who hates [to sit watching television] might **instead** [try skydiving].

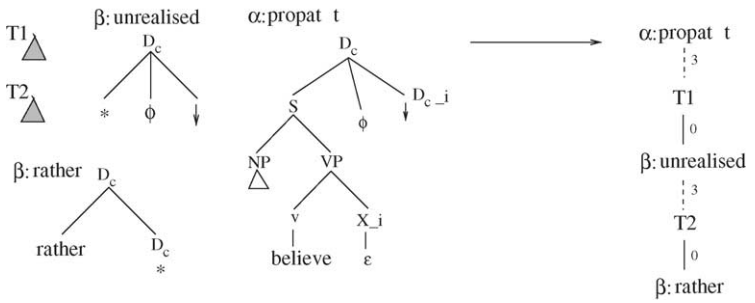


Fig. 16. D-LTAG derivation of Example 17—"Epigenesists believed that the organism was not yet formed in the fertilized egg. Rather, it arose as a consequences of profound changes in shape and form during the course of embryogenesis."

Empirical evidence for this distinction between structural connectives and discourse adverbials comes from Creswell et al. (2002), who describe an annotation experiment in which annotators were asked to identify the minimal text unit in the preceding discourse containing the source of the "left-hand" argument of the following nine connectives:

- Resultatives: *as a result, so, therefore*
- Additives: *also, in addition, moreover*
- Concessives: *nevertheless, yet, whereas*

The data came from Brown corpus, WSJ corpus, Switchboard corpus, and 58 transcribed oral histories of online Social Security Administration (SSA) Oral History Archives.[10] The results showed a variety of distribution patterns:

- *So* always took the immediately preceding sentence or sequence of sentences as its left argument.
- *Nevertheless* often took XP (i.e., phrasal) arguments.
- *Therefore* often took its left-hand argument from a subordinate clause.

Connectives that patterned with *so* were taken to be structural connectives, while the others were taken to get their "left-hand" argument anaphorically from inter alia a non-adjacent clause, a relative clause, etc., that is, from a clause that is not structurally connected to the discourse adverbial.

The problems of dealing with these two types of discourse connective differ. With structural connectives, one has to rely on the parser to associate a connective with its intended arguments: An incorrect attachment decision will mean an incorrect argument assignment. With anaphoric connectives, as with any anaphor, one must develop a procedure for resolving them.

Now it is well-known that different anaphors display different patterns vis-a-vis the distribution and type of their antecedents: plural pronouns allow *split antecedents* while singular pronouns do not; definite noun phrases (NPs) commonly allow antecedents related through *bridging* while pronouns do so only rarely; the antecedents of demonstrative pronouns commonly derive from clauses, while those of personal pronouns most commonly derive from NPs; etc. In the case of a discourse adverbial, if its "left-hand" argument is anaphoric, then one needs to articulate a procedure for finding its antecedent and from that, deriving its argument.

We do not think all discourse adverbials will pattern exactly the same vis-a-vis their antecedents, so we are proceeding on a case-by-case basis to gather data on how they pattern and on what features are relevant to that patterning. The preliminary study we have carried out on the discourse adverbial *instead*Miltsakaki et al. (2003) illustrates what is needed. Here I will summarize and elaborate on that study and comment on how we are now proceeding. *Instead* comes in two forms: (i) a bare adverbial, as in

(23) *Instead* John ate an apple.

and (ii) modified by an "of" PP, as in

(24) John ate an apple *instead of* a pear.

(25) John spent the afternoon at the zoo *instead of* at the museum.

With an "of" PP, both args of *instead* derive *structurally*: the first from the modified phrase (e.g., "an apple") and the second from the "of" PP (e.g., "a pear"). Semantically, that second argument is a salient but unchosen *alternative* to the first, with respect to the given predication. This is basic to the interpretation of *instead* in both its modified and bare forms.

As a bare adverbial, *instead* continues to get its first argument structurally, but its second argument – the salient but unchosen alternative – must be derived *anaphorically*, from the discourse context. But not every context provides alternatives:

(26) a. John found it hard to eat an apple. *Instead* he ate a pear.
b. John found it easy to eat an apple. #*Instead* he ate a pear.
c. I {told, expected} John to eat an apple. *Instead* he ate a pear.
d. John {told, expected} me to eat an apple. #*Instead* he ate a pear.

As far as I am aware, there is no theoretical account of what types of phrases/clauses suggest alternatives that license "instead".[11]

To begin to discover this empirically, pairs of annotators separately examined 100 successive instances of bare *instead* in the Penn TreeBank and recorded the minimal text span containing the antecedent of its anaphoric argument. There was agreement in 97/100 cases, and the other three cases were excluded from further analysis.

We then chose features to annotate that we had observed in serendipitously encountered instances of *instead*:

- clausal negation

(27) John *couldn't* sleep. Instead, he wrote code. (**Verbal neg**)
(28) *No one* could sleep. Instead, everyone wrote code. (**Subj neg**)
(29) John ate *none of his spinach*. Instead, he fed it to his frog. (**Obj neg**)

- presence of a monotone-decreasing quantifier (**MDQ**)

(30) *Few* students like to do homework. Instead, they would rather party.
(31) Students *seldom* sleep in class. Instead, they take notes assiduously.

- presence of a modal auxiliary (**Modal**)

(32) You *should* exercise more. Instead you sit like a couch potato.

- whether the antecedent is embedded in a higher clause (**Embed**)

(33) *John wanted* to eat a pear. *Instead*, he ate an apple.
(34) *Chrysler officials resisted* cutting output. *Instead*, they slapped $1000 cash rebates on vehicles.
(35) *Paine Webber considered* recommending specific stocks. *Instead*, it just urged its clients to stay in the market.

The results are shown in Fig. 17.[12]

We then investigated whether other clauses that don't serve as antecedents for *instead*, which we call "potentially competing antecedents" or "PCAs", have a similar distribution with respect to these features. As in Soon, Ng, and Lim (2001), we limited potentially competing antecedents to ones occurring between the anaphor and its true antecedent. Here, PCAs were finite or non-finite clauses intervening between *instead* and its true antecedent. For the 97

Features	YES (of 97)	NO (of 97)
Verbal neg	37 (38%)	60 (62%)
Subj neg	5 (5%)	92 (95%)
Obj neg	10 (10%)	82 (85%)
MDQ	1 (1%)	96 (99%)
Modal	12 (12%)	85 (88%)
Condit	1 (1%)	96 (99%)
Embed	57 (59%)	40 (41%)

Fig. 17. Distribution of features of the antecedent of *instead*.

	Antecedents		PCAs	
Features	YES (of 97)	NO (of 97)	YES (of 169)	No (of 169)
Verbal neg	37 (38%)	60 (62%)	21 (12%)	148 (88%)
Subj neg	5 (5%)	92 (95%)	8 (5%)	161 (95%)
Obj neg	10 (10%)	82 (85%)	6 (4%)	139 (82%)
MDQ	1 (1%)	96 (99%)	0 (0%)	169 (100%)
Modal	12 (12%)	85 (88%)	17 (10%)	152 (90%)
Condit	1 (1%)	96 (99%)	0 (0%)	169 (100%)
Embed	57 (59%)	40 (41%)	14 (8%)	155 (91%)

Fig. 18. Distribution of features of the PCAs of *instead*.

tokens of *instead* on which annotators agreed, this produced 169 PCAs. The distribution of the same seven features for these PCAs is shown in Fig. 18.

There are some obvious differences between the antecedents and PCAs of *instead*. First, as shown in the following summary of clausal negation features

	Antecedents		PCAs	
Features	YES (of 97)	NO (of 97)	YES (of 169)	NO (of 169)
Verbal neg	37 (38%)		21 (12%)	
Subj neg	5 (5%)		8 (5%)	
Obj neg	10 (10%)		6 (4%)	

clausal negation was found to be over 2.5 times more common in the antecedent of *instead* than in PCAs—52/97 times (≈53%) versus 35/169 times (≈20%).

Second, focusing on the **embed** feature

	Antecedents		PCAs	
Features	YES (of 97)	NO (of 97)	YES (of 169)	NO (of 169)
Embed	57 (59%)		14 (8%)	

the antecedent of the anaphoric argument of *instead* was found to be over seven times more frequently embedded in a higher verb than a PCA was—57/97 times (≈59%) versus 14/169 times (≈8%).

On the other hand, for the features related to the antecedent being in a conditional (**condit**) or containing a monotonically decreasing quantifier (**MDQ**), there isn't enough data to draw

any conclusions. The feature related to the antecedent containing a modal auxiliary (**Modal**) does not, as such, seem at all predictive.

Subsequent to this study, we reviewed the data and decided that this initial feature set should be refined in at least the following ways, to widen the difference between antecedents and PCAs.

1. Although the embedding feature is strongly predictive, we realized that not all embedding contexts suggest alternatives to their embedded clauses. In particular, some embedded PCAs (but no embedded antecedents of *instead*) were embedded under factive verbs like *know*. It is well-known that factive verbs presuppose the truth of their embedded clause (Kiparsky & Kiparsky, 1970), as in

(36) John knows that Fred eats meat.

They therefore do not provide alternatives that can serve as antecedents for *instead*, cf.

(37) John believes/*knows that Fred eats meat. Instead Fred eats tofu.

Therefore, we should annotate a feature on the embedding verb, identifying whether or not it is factive, to exclude clauses embedded under the latter as potential antecedents. Since there is only a small number of factive verbs (although they are relatively common), such a feature could be annotated automatically, with high reliability.

2. Certain verbs appear to suggest alternatives, independent of whether the clause also contains explicit negation, a monotonically-decreasing quantifier, a modal auxiliary or clausal embedding. Consider the following examples.

(38) John *doubted* Mary's resolve. *Instead*, he thought she would give up as soon as he left.

(39) NBC is contemplating *getting out of* the cartoon business. *Instead*, it may "counter-program" with shows for an audience that is virtually ignored in that time period: adults.

(40) Investors have *lost* their enthusiasm for the stock market. *Instead*, they are buying government bonds.

(41) But respectability still *eludes* Italy's politics. *Instead*, it has the phenomenon of Mr. Berlusconi.

Many additional such verbs have come to our attention. They appear to fall roughly into two classes, although neither corresponds to any known thesaurus or WordNet class. The first class – including *doubt*, *refuse*, *deny*, *preclude*, etc. – appears to contain an element of implicit negation, and might be called *negative propositional attitude verbs*. The second class – including *stop*, *lose*, *get out of*, *change*, *drop*, *give up*, *elude*, etc. – might be called *n*egative state change verbs. They indicate that in the situation after the event conveyed by the clause, some earlier feature of the situation no longer holds. This feature then seems to be available as an alternative to the indicated change.

While verbs in both classes appear to suggest alternatives, the composition of these classes remains to be specified. So we must acquire their membership concurrently with carrying out annotation.

3. Even more of a challenge to automatic identification, is the fact that other lexico-syntactic elements that do not fall into a priori classes appear able to suggest alternatives as well. In the following example from the Penn TreeBank

(42) The tension was evident on Wednesday evening during Mr. Nixon's final banquet toast, normally an opportunity for reciting platitudes about eternal friendship. *Instead*, Mr. Nixon reminded his host, Chinese President Yang Shangkun, that Americans haven't forgiven China's leaders for the military assault of June 3–4 that killed hundreds, and perhaps thousands, of demonstrators.

either the adverb "normally" or the noun "opportunity" appears to be a sufficient trigger for alternatives and hence the use of *instead*:

(43) Normally, we eat pasta on Tuesday. *Instead*, tonight we're having fish.
(44) John had the opportunity to buy a cheap used car. *Instead*, he bought a scooter.

So while it is clear that we should broaden the range of features being considered, it is not clear how to go about identifying them, except by noticing them in the context of *instead*.

Finally, I should comment on *relational features* that derive from the *pair* of structural and anaphoric arguments to *instead*—for example, whether the two have the same surface subject (as in most, but not all, of the examples above), or related subjects, as in Example 45.

(45) In an abrupt reversal, the United States and Britain have indefinitely put off their plan to allow Iraqi opposition forces to form a national assembly and an interim government by the end of the month. *Instead*, top American and British diplomats leading reconstruction efforts here told exile leaders in a meeting tonight that allied officials would remain in charge of Iraq for an indefinite period, said Iraqis who attended the meeting.

While *relational features* appear relevant to resolving *instead*, they were not included in our original feature set. But it is clear that relational features should be included as well. The context in which we will examine these and other features is the Penn Discourse TreeBank.

5. Penn Discourse TreeBank

The Penn Discourse TreeBank (http://www.cis.upenn.edu/∼pdtb) aims to do for discourse what the Penn TreeBank has done for sentence-level processing, that is, to provide a sharable resource for the development of automated techniques of discourse analysis and generation. The value of a TreeBank comes from the "knowledge" added to it, over and beyond its sequence of sentences. When complete and released (around November 2005), it is expected to have approximately 20,000 annotations of the 250 types of explicit connectives identified in the corpus, and 10,000 annotations of *implicit conenctives* (see below).

Creating the Penn Discourse TreeBank (PDTB) involves manually identifying, annotating and assessing inter-annotator agreement on (a) all discourse connectives in the Penn TreeBank and (b) the text segments from which each connective draws its arguments (Miltsakaki et al., 2004). While the PDTB reflects the theoretical bias of D-LTAG in terms of a lexical basis for discourse analysis and different types of discourse connectives, the instructions to annotators[13] only require them to identify the minimal spans of text whose meaning is involved in the use of a particular connective. These spans may cover inter alia an embedded clause, as in the first (anaphoric) argument to *instead* in Example 46, a previous (non-adjacent) clause, as in the first (anaphoric) argument to *o*therwise in Example 47, or the immediately preceding sentence or clause, as in Example 48.

(46) Anne Compoccia wanted [to be a nun].
Instead, [she found herself in prison for embezzling city funds].

(47) [If the light is red], stop.
Otherwise, [just continue down the road.]

(48) [There are no separate rafters in a flat roof];
instead, [the ceiling joists of the top story support the roofing.]

(Other possibilities include the immediately preceding discourse, a string that doesn't correspond to an existing syntactic constituent, or even a discontinuous string.)

PDTB annotation is produced using WordFreak,[14] an annotation tool developed by Tom Morton and then modified by Jeremy Lacivita to satisfy the needs of PDTB annotation. To support multi-level analysis, annotation is rendered in XML as "stand-off" annotation, aligned with similar stand-off versions of the Penn TreeBank syntactic annotation and the predicate–argument annotation of PropBank (Kingsbury & Palmer, 2002). In the first tranch of connectives to be annotated were the discourse adverbials *instead*, *otherwise*, *nevertheless*, *as a result* and *therefore*, and the subordinate conjunctions *because* (both alone and when preceded by *partly*, *in part*, *only*, *just* or *largely*), *although*, *even though*, *when* (both alone and when preceded by *just*, *only*, *even* or *largely*) and *so that*.

In addition, the PDTB is annotating *implicit connectives* between adjacent sections with no explicit connective between them. Here, the two sentences are taken to be the two arguments, and the annotators are asked to provide, where possible, an explicit connective that captures the inferred relation between them. For example,

(49) [The 6 billion that some 40 companies are looking to raise in the year ending March 31 compares with only 2.7 billion raised on the capital market in the previous fiscal year]. *IMPLICIT*-(In contrast) [In fiscal 1984 before Mr. Gandhi came to power, only 810 million was raised].

The final version of the PDTB will also contain characterizations of the semantic roles associated with the arguments of each type of connective, similar to both PropBank annotation of the semantic roles of verbs (Kingsbury & Palmer, 2002) and NomBank annotation of the semantic roles of nouns (Meyers et al., 2004). Such role annotations will allow software running over the PDTB to distinguish between different senses of a connective (e.g., temporal versus

concessive *while*) or, for example, to back off to all connectives that share the same set of semantic roles.

Further discussion of the PDTB, its annotation guidelines and levels of inter-annotator agreement can be found in Miltsakaki et al. (2004) and Prasad et al. (2004).

The Penn Discourse TreeBank is not the first or only effort to annotate discourse structure. Efforts to do so started over 10 years ago, as a way of providing empirical justification for high-level theories of discourse structure (Grosz & Sidner, 1986; Moser & Moore, 1996). Although much time and energy was devoted to the work (Di Eugenio, Jordan, Moore, & Thomason, 1998), the results have not been widely used in the computational arena, unlike the Penn TreeBank. It is hoped that current efforts will not suffer this fate.

The work closest to the Penn Discourse TreeBank in English is the corpus developed by Carlson and Marcu and their colleagues (Carlson, Marcu, & Okurowski, 2002; Marcu, 1999) based on Rhetorical Structure Theory (Mann & Thompson, 1988).[15] RST is a theory of discourse analysis that holds that (1) there is a specified set of rhetorical relations that can hold between adjacent units of discourse; (2) adjacent units of discourse are related by a single rhetorical relation that accounts for the semantic or pragmatic (intentional) sense associated with their adjacency; (3) units so related form larger units that participate in rhetorical relations with units that they themselves are adjacent to; and (4) in many, but not all, such juxtapositions, one of the units (the satellite) provides support for the other (the nucleus), which then appears to be the basis for rhetorical relations that the larger unit participates in. Given these principles, the main aspects of RST annotation are (1) demarcating the elementary discourse units that underpin the representation; (2) identifying how they fit together into larger spans; and (3) annotating the particular rhetorical relation that holds between elements that form a larger span.

The RST-annotated corpus[16] differs from the Penn Discourse TreeBank in several ways—the most significant being the difference in theoretical perspective. The RST-corpus is based on an a priori set of rhetorical relations, and annotators are given specific instructions as to when each should be chosen as the annotation for a text. In contrast, the PDTB is grounded in the corpus itself: While annotators may be instructed as to when to consider a particular token a discourse connective (as opposed to, e.g., a wh-complementizer or a relative pronoun), once a token is judged to be a connective, the annotators' job is to identify its two arguments in the corpus. Operationally, this means that RST annotation starts with identifying discourse units and then selecting what rhetorical relations holds between them, while PDTB starts with identifying connectives and then what it is that they connect.

We are not downplaying the importance of having an annotated corpus of coherence relations associated with adjacent discourse units. But we believe that the task of producing such a corpus can be made easier by having already identified the higher order predicate–argument relations associated with explicit discourse connectives. They can then be factored into the calculation or removed from the calculation, as appropriate (Webber et al., 2003).

6. Conclusion

This paper has reviewed our work on a lexicalized grammar for low-level discourse, explaining what has motivated the work and what it achieves, including

- allowing us to make specific generalizations about how lexico-syntactic elements contribute to the syntax and semantics of both the clause and discourse, and how those contributions may interact.
- opening up the (still to be realized) possibility of allowing sentence processing and low-level discourse processing to be integrated.
- allowing us to develop a large, reliably annotated corpus in which the basis for annotation decisions – discourse connectives (viewed as predicates) and their arguments – is clear.

For the next few years, the Penn Discourse TreeBank is the future of D-LTAG. It will provide a Gold Standard for further parser development for D-LTAG, and through its integration with the Penn TreeBank and PropBank, enable the development of data-intensive, probabilistic methods for resolving anaphoric connectives. It will undoubtedly be a source of interesting data and interesting ideas for many years to come.

Notes

1. Other *lexicalized grammars* include Combinatory Categorial Grammar (Steedman, 1996) and Dependency Grammar (Melcuk, 1988). Lexicalized grammars have proved to be a significant tool in the theoretical understanding of clause-level phenomena and have spurred computational development of robust, wide-coverage parsers for Natural Language text (Bangalore & Joshi, 1999; Clark & Hockenmaier, 2002; Hockenmaier & Steedman, 2002).
2. I thank one of the Cognitive Science reviewers for pointing this out.
3. We are only beginning to explore this aspect of LTAG now at the discourse level. In LTAG, each node in a tree has an associated *feature structure* that can, along with the node label, be used to constrain possible substitutions and/or adjunctions at that node. While such feature structures are not discussed in this paper, see Forbes-Riley et al. (2004).
4. This simple recursion is related to *dominant topic chaining* in Scha and Polanyi (1988) and *entity chains* in Knott, Oberlander, O'Donnell, and Mellish (2001). But null connectives are also compatible with the inference that a stronger relation (such as *explanation*) holds between discourse clauses. If such an inference does hold, then it would no longer be a case of *dominant topic chaining* or *entity chains*.
5. Syntactically, LTAG doesn't distinguish between discourse adverbials such as *instead* and clausal adverbials such as *swiftly*, *annually* or *unfortunately*. They are all associated with the same set of auxiliary trees because they can all appear at the same positions within the clause. Forbes (2003) gives an extensive analysis of the features of an adverbial that lead it to be interpreted as a discourse adverbial rather than a clausal modifier.
6. A separate version of the discourse parser uses LexTract (Xia, Palmer, & Joshi, 2000) at the sentence-level and LEM at the discourse-level. LexTract provides unique TAG derivations for sentences in the Penn TreeBank, so that heuristics are not needed to select trees or choose attachment points. This just avoids sever ambiguity problems at

the sentence-level, in order to focus on discourse-level processing. The process following the use of LexTract to produce unique sentence-level derivations is the same as in Fig. 9.

7. While one does indeed want to identify, for Information Structure, where a clause-medial adverbial occurs in clause structure, doing it via this copy-and-replace mechanism is specific to this particular implementation. A process that interleaved clausal parsing with discourse parsing would, presumably, identify a medial adverbial where it occurs and process it at that point.
8. Rashmi Prasad, personal communication.
9. Following the conventions used in the Penn Discourse TreeBank (Section 5), arguments are bracketed, while connectives are underlined and in bold.
10. http://www.ssa.gov/history/orallist.html.
11. Forbes (2003) shows that these are not the same alternatives that underpin the semantics of focus particles such as "only" and "even". On the other hand, there are clearly relationships between them, as "Only John ate an apple. *Instead* the other boys ate pears."
12. Antecedents could display one or more compatible features, e.g., both **Subj neg** and **Modal**.
13. http://www.cis.upenn.edu/∼pdtb/manual/pdtb-tutorial.pdf.
14. http://www.sourceforge.net/projects/wordfreak.
15. For German, there is now a similar effort to annotate discourse connectives as part of the Potsdam Commentary Corpus (Stede, 2004).
16. Distributed now by the Linguistic Data Consortium, http://www.ldc.upenn.edu.

Acknowledgements

Much of the material in this article derives from talks I have given and from papers co-authored by members of the D-LTAG group—Cassandre Creswell, Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad and Aravind Joshi, at the University of Pennsylvania, and myself, at the University of Edinburgh. The paper has also gained from on-going discussions among group members, in connection with the development of the Penn Discourse TreeBank, from comments from Mark Steedman, and from suggestions from Fernanda Ferreira and two other (anonymous) *Cognitive Science* reviewers. The Penn Discourse TreeBank project is partially supported by NSF Grant EIA 0224417 (Joshi).

References

Asher, N. (1993). *Reference to abstract objects in discourse*. Boston, MA: Kluwer.

Asher, N., & Lascarides, A. (1998). The semantics and pragmatics of presupposition. *Journal of Semantics*, *15*(3), 239–300.

Bangalore, S., & Joshi, A. (1999). Supertagging: An approach to almost parsing. *Computational Linguistics*, *25*(2), 237–265.

Carlson, L., Marcu, D., & Okurowski, M. E. (2002). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the second SIGdial workshop on discourse and dialogue*.

Clark, S., & Hockenmaier, J. (2002). Evaluating a wide-coverage CCG parser. In *Proceedings of the LREC workshop on beyond Parseval* (pp. 60–66).

Creswell, C., Forbes, K., Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2002). The discourse anaphoric properties of connectives. In *Proceedings of the discourse anaphora and anaphor resolution colloquium*.

Cristea, D., & Webber, B. (1997). Expectations in incremental discourse processing. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics (ACL97/EACL97)* (pp. 88–95).

Di Eugenio, B., Jordan, P. W., Moore, J. D., & Thomason, R. H. (1998). An empirical investigation of proposals in collaborative dialogues. In *Proceedings of COLING/ACL'98* (pp. 325–329).

Forbes, K. (2003). *Discourse semantics of s-modifying adverbials*. Ph.D. thesis, Department of Linguistics, University of Pennsylvania.

Forbes, K., Miltsakaki, E., Prasad, R., Sarkar, A., Joshi, A., & Webber, B. (2001). D-LTAG system—Discourse parsing with a lexicalized tree-adjoining grammar. In *ESSLLI'2001 workshop on information structure, discourse structure and discourse semantics*.

Forbes, K., & Webber, B. (2002). A semantic account of adverbials as discourse connectives. In *Proceedings of third SIGDial workshop* (pp. 27–36).

Forbes-Riley, K., Webber, B., & Joshi, A. (in press). Computing discourse semantics: The predicate–argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*.

Gardent, C. (1997). *Discourse tree adjoining grammars*. Claus report no. 89, University of the Saarland, Saarbrücken.

Grosz, B., & Sidner, C. (1986). Attention, intention and the structure of discourse. *Computational Linguistics*, *12*(3), 175–204.

Hockenmaier, J., & Steedman, M. (2002). Generative models for statistical parsing with combinatory categorial grammar. In *Proceedings of 40th annual meeting of the Association for Computational Linguistics*.

Hovy, E. (1988). *Generating natural language under pragmatic constraints*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Joshi, A. (1987). An introduction to Tree Adjoining Grammar. In A. Manaster-Ramer (Ed.), *Mathematics of language* (pp. 87–114). Amsterdam: John Benjamins.

Kingsbury, P., & Palmer, M. (2002). From TreeBank to PropBank. In *Proceedings of the third international conference on language resources and evalution (LREC)*.

Kiparsky, P., & Kiparsky, C. (1970). Fact. In M. Bierwisch & K. E. Heidolph (Eds.), *Progress in linguistics* (pp. 143–173). Mouton.

Knott, A., Oberlander, J., O'Donnell, M., & Mellish, C. (2001). Beyond elaboration: The interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 181–196). John Benjamins.

Linde, C. (1974). *The linguistic encoding of spatial information*. Ph.D. thesis, Department of Linguistics, Columbia University.

Mann, W., & Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*(3), 243–281.

Marcu, D. (1999). *Instructions for manually annotating the discourse structure of texts* Available at http://www.isi.edu/marcu/software/manual.ps.gz. Accessed on 28 May 2004.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. MIT Press.

McKeown, K. (1985). *Text generation: Using discourse strategies and focus constraints to generate natural language texts*. Cambridge, UK: Cambridge University Press.

Melcuk, I. (1988). *Dependency syntax: Theory and practice*. Albany, NY: State University of New York.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zilinska, V., Young, B., et al. (2004). The NomBank project: An interim report. In *NAACL/HLT workshop on frontiers in corpus annotation*.

Miltsakaki, E., Creswell, C., Forbes, K., Joshi, A., & Webber, B. (2003). Anaphoric arguments of discourse connectives: Semantic properties of antecedents versus non-antecedents. In *EACL workshop on computational treatment of anaphora*.

Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2003). Annotating discourse connectives and their arguments. In *NAACL/HLT workshop on frontiers of corpus annotation*.

Moore, J. (1990). *Participating in explanatory dialogues*. Boston, MA: MIT Press.

Moser, M., & Moore, J. (1996). Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, *22*(3), 409–419.

Polanyi, L., & van den Berg, M. H. (1996). Discourse structure and discourse interpretation. In P. Dekker & M. Stokhof (Eds.), *Proceedings of the 10th Amsterdam colloquium* (pp. 113–131). University of Amsterdam.

Prasad R., Miltsakaki, E., Joshi, A., & Webber, B. (2004). Annotation and data mining of the Penn Discourse TreeBank. In *Proceedings of ACL/EACL Workshop on Discourse Annotation* (pp. 88–95).

Quirk, R., Greenbaum, S., Leech, G., & Svartik, J. (1972). *A grammar of contemporary English*. Harlow: Longman.

Sarkar, A. (2000). Practical experiments in parsing using tree-adjoining grammars. In *Proceedings of the fifth TAG+ workshop* (pp. 193–198).

Scha, R., & Polanyi, L. (1988). An augmented context free grammar for discourse. In *Proceedings of the 12th international conference on computational linguistics (COLING'88)* (pp. 573–577).

Schabes, Y. (1990). *Mathematical and computational aspects of lexicalized grammars*. Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.

Schilder, F. (1997). Tree discourse grammar, or how to get attached to a discourse. In *Proceedings of the second international workshop on computational semantics*.

Sibun, P. (1992). Generating text without trees. *Computational Intelligence*, *8*(1), 102–122.

Soon, W. M., Ng, H. T., & Lim, D. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, *27*(4), 521–544.

Stede, M. (2004, July). The Potsdam commentary corpus. In *ACL workshop on discourse annotation*.

Steedman, M. (1996). *Surface structure and interpretation*. Cambridge, MA: MIT Press.

Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic Inquiry*, *34*, 649–689.

Webber, B. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, *6*(2), 107–135.

Webber, B., & Joshi, A. (1998). Anchoring a lexicalized tree-adjoining grammar for discourse. In *Coling/ACL workshop on discourse relations and discourse markers* (pp. 86–92).

Webber, B., Joshi, A., & Knott, A. (2000). The anaphoric nature of certain discourse connectives. In *Making sense: From Lexeme to discourse*.

Webber, B., Knott, A., & Joshi, A. (2001). Multiple discourse connectives in a lexicalized grammar for discourse. In H. Bunt, R. Muskens, & E. Thijsse (Eds.), *Computing meaning: Vol. 2* (pp. 229–249). Kluwer.

Webber, B., Knott, A., Stone, M., & Joshi, A. (1999). Discourse relations: A structural and presuppositional account using lexicalized TAG. In *Proceedings of the 36th annual meeting of the Association for Computational Linguistics* (pp. 41–48).

Webber, B., Knott, A., Stone, M., & Joshi, A. (1999). What are little trees made of: A structural and presuppositional account using lexicalized TAG. In *Proceedings of international workshop on levels of representation in discourse (LORID'99)* (pp. 151–156).

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

Xia, F., Palmer, M., & Joshi, A. (2000). A uniform method of grammar extraction and its applications. In *Proceedings of the SIGDAT conference on empirical methods in natural language processing (EMNLP)*.

XTAG-Group. (2001). *A lexicalized tree adjoining grammar for English*Tech. Rep. IRCS 01–03, University of Pennsylvania, available at ftp://ftp.cis.upenn.edu/pub/ircs/technical-reports/01–03, accessed on 28 May 2004.