Sentential Structure and Discourse Parsing

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, David Ahn[1]
FX Palo Alto Laboratory
3400 Hillview Ave, Bldg. 4
Palo Alto, CA 94304
{polanyi|culy|vdberg|thione}@fxpal.com, ahn@science.uva.nl

Abstract

In this paper, we describe how the LIDAS System (Linguistic Discourse Analysis System), a discourse parser built as an implementation of the Unified Linguistic Discourse Model (U-LDM) uses information from sentential syntax and semantics along with lexical semantic information to build the Open Right Discourse Parse Tree (DPT) that serves as a representation of the structure of the discourse (Polanyi et al., 2004; Thione 2004a,b). More specifically, we discuss how discourse segmentation, sentence-level discourse parsing, and text-level discourse parsing depend on the relationship between sentential syntax and discourse. Specific discourse rules that use syntactic information are used to identify possible attachment points and attachment relations for each Basic Discourse Unit to the DPT.

1 Introduction

In this paper, we describe discourse parsing under the Unified Linguistic Discourse Model (U-LDM) (Polanyi et al. 2004). In particular, we describe the relationship between the output of sentential parsing and discourse processing.

The goal of discourse parsing under the U-LDM is to assign a proper semantic interpretation to every utterance in a text. In order to do so, the model constructs a structural representation of relations among the discourse segments that constitute a text. This representation is realized as a Discourse Parse Tree (DPT). Incoming discourse utterances are matched with the informational context needed for interpretation, and attached to nodes on the right edge of the tree.

1.1 Discourse Parse Tree

The U-LDM builds upon the insights and mechanisms of the Linguistic Discourse Model (LDM) (Polanyi 1988). The DPT specifies which segments are coordinated to one another (bear a similar relationship to a common higher order construct), which are subordinated to other constituents (give more information about entities or situations described in that constituent, or, alternatively, interrupt the flow of the discourse to interject unrelated information), and which are related as constituents of logical, language specific, rhetorical, genre specific or interactional structures (n-ary relations). Importantly, the representation identifies which constituents are available for continuation at any moment in the development of a text and which are no longer structurally accessible.

1.2 Discourse Parsing

While full understanding of the meaning of constituent utterances, world knowledge, inference and complex reasoning are needed to create the correct Discourse Parse Tree to represent the structure of discourse under the LDM, in developing the U-LDM, it became apparent that most of the information needed to assign structural relations to a text can be recovered from relating lexical, syntactic and semantic information in constituent sentences to information available at nodes on the DPT. Largely formal methods involving manipulating information in lexical ontologies and the output of sentential syntactic and semantic analysis are sufficient to account for most cases of discourse continuity, and give rise to only limited ambiguity in most other cases.[2] The assignment of correct temporal, personal and spatial interpretation to utterances, which relies in large part on the relative location of referential expression and their

[1] Current address:
Language and Inference Technology Group,
Informatics Institute
Kruislaan 403
1098 SJ Amsterdam, The Netherlands

[2] Complex default logic based reasoning as in Structured Discourse Representation Theory (Asher 1993; Asher and Lascarides 2003), speculations about the intentions or beliefs of speakers (as in Grosz and Sidner (1986)), or the intricate node labeling exercises familiar from *Rhetorical Structure Theory* (Mann and Thompson 1988; Marcu 1999, 2000) are not necessary.

referents in the DPT representation of the structure of the discourse, can then often be recovered. From a computational point of view, this is good news.

Parsing consists of the following steps for every sentence of the discourse.

1. The sentence is parsed by a sentence level parser, in our case the LFG-based Xerox Linguistic Environment (XLE).
2. The sentence is broken up in discourse relevant segment based on the syntactic information from the sentence parser.
3. The segments are recombined into one or more small discourse trees, called Basic Discourse Unit (BDU) trees, representing the discourse structure of the sentence.
4. The BDU trees corresponding to the sentence are each attached to the DPT tree.[3]

In the remainder of this paper, we will describe how the LIDAS System (Linguistic Discourse Analysis System), a discourse parser built as an implementation of the U-LDM, uses information from sentential syntax and semantics along with lexical semantic information to build up the structural representation of source texts[4]. Specifically, we will discuss discourse segmentation, sentence-level discourse parsing, and text-level discourse parsing—phases of the discourse parsing process that depend on the relationship between sentential syntax and discourse.

2 Discourse Segments and Basic Discourse Units

U-LDM discourse segmentation is based on the syntactic reflexes of the semantic content of the linguistic phenomena making up discourse. Since elementary discourse units must be identified to build up discourse structure recursively, *discourse segments* under the U-LDM are identified as the syntactic units that encode a *minimum unit* of *discourse function* and/or *meaning* that must be interpreted relative to a set of contexts to be understood. Minimal Functional units include greetings, connectives, discourse PUSH/POP markers and other "cue phrases" that connect or modify content segments. Minimal meaning units are units that express information about not more than one event, event-type or state of affairs in a possible world. Roughly speaking they are "elementary propositions" or "event-type predicates" corresponding in (neo-) Davidsonian semantics to an elementary statement that contains at most one event-quantifier. Structurally, a minimum meaning unit does not contain a proper subpart that itself communicates content and has a syntactic correlate that can stand on its own.

Under the U-LDM, segments can be *discontinuous* (if there is overt material on both sides of an intervening segment) or *fragmentary*. A single word answer to a question is a complete segment, whereas the same word uttered in an incomplete and unrecoverable phrase is a fragment.

The U-LDM defines segmentation in purely sentence syntactic terms. However, the choices of definitions are motivated by semantic considerations. For example, since auxiliaries and modals do not refer to events distinctly from the main verb, they do not form segments separate from the corresponding main verbs. By the same reasoning, other modal constructions that involve infinitives (e.g. *have to*, *ought to*, etc.) also constitute a single segment with their complements as do Cleft constructions, despite the presence of two verbs.[7] On the other hand, Equi verbs (e.g., *try*, *persuade*) and Raising verbs (e.g., *seem*, *believe*, etc) form separate BDUs from their verbal complements, since both eventualities can be continued. In contrast, even though event nominals, including gerunds, refer to eventualities (possibly) distinct from the verbs of which they are arguments or adjuncts,

Content segments (BDUs)	Simple clauses, Subordinate clauses and participial phrases, Secondary predications,[5] Interpolations (e.g. parentheticals, appositives, interruptions, etc.), Fragments (e.g. section headings, bullet items, etc.)
Operator segments	Conjunctions that conjoin segments, Discourse operators (e.g. "scene-setting" preposed modifiers)[6]Discourse connectives

Table 1: Examples BDU and Operator Segments.

[3] If a sentence is sufficiently complex, it may consist of two or more completely different independent discourse units. Such cases are treated as if the two parts of the sentence were really two different sentences. One consequence of this is that a syntactic coordination of two sentences may correspond to a subordination in the DPT, because they are treated independently.

[4] The LiveTree environment for discourse parsing is described in detail in Thione 2004b.

[5] For example, "Chris served the chapter | as Social Chairman"

[6] For example: "Finally, | the audio is summarized."

[7] For example: "It **is** segmentation that we **are discussing**."

those eventualities can not (easily) be continued and therefore are not segments.[8]

The U-LDM, in contrast to other discourse theories, has a fine-grained taxonomy of minimal units. The most prominent distinction of discourse segments is between *Basic Discourse Units (BDUs*) and *Operator segments* (see Table 1). BDUs are segments realized in a linguistic utterance that can independently provide the context for segments later in the discourse. Operator segments can not do so, they can only provide context for segments in their scope.

In the following section, we explain more fully how sentence syntax is used to segment sentences into discourse segments.

2.1 The role of the XLE in discourse segmentation

Dividing a text into discourse segments begins by applying a sentence breaking algorithm to determine the tokens to be segmented. These tokens are normally complete sentences, but may also be fragments in some cases such as titles or elliptical phrases such as "Yes". The tokens are processed by a *discourse segmenter*. After the segmenter has completed its work, segments are passed to a U-LDM *BDU-Tree parser*, which constructs one or more BDU-Trees from the BDUs identified in each sentence.

The LIDAS segmenter first sends the input chunk to be parsed by the Xerox Linguistic Environment parser (XLE) (Maxwell and Kaplan 1989). The XLE is a robust Lexical Functional Grammar (LFG) parser. The XLE tries to parse the input, and returns either a packed representation of all possible parses or the most probable parse as selected by its stochastic disambiguation component (Riezler et al. 2002). If the XLE can not find a parse of the input as a complete sentence it tries to construct a sequence of partially parsed fragments.[9] The LIDAS segmenter segments the most probable parse and, as a backup, the first parse from the packed representation, if the first and the most probable parse differ.

[8] Note that the contrast between modal verbs on the one hand and Raising and Equi verbs on the other clearly shows that the *surface form* of a phrase (e.g., the infinitival complements) is not always sufficient to determine segment status. Similarly, a finite verb is not a sufficient indicator of a segment, as seen in the cleft constructions. Crucially, and appropriately, we need to refer to the discourse property of potential continuation.

[9] The XLE has a failure rate (i.e., no parse whatsoever) of approximately 0.5% on our corpus of technical reports.

Content Segments
Certain F-structures with subjects[10]
Fragments
Parentheticals
Headers
Syntactic coordination (except conjunct itself)
Operator segments
Conjuncts in coordination
Initial comma separated modifiers
Subordinating conjunctions

Table 2. Segment classification configurations.

The parse information consists of a **c**(onstituent) structure (essentially a standard parse tree) and a **f**(unctional) structure containing predicate-argument information, modification information, and other grammatical information such as tense and agreement features. F-structures make up the primary source of linguistic information used in discourse parsing.

To identify the discourse segments within the sentence, seven syntactic configurations in c-structure and f-structure are examined. The relatively small set of configurations in Table 2 accounts for the full range of discourse segments.

According to these segmentation rules, it is possible for a segment to be embedded in another segment. Because on the tree-projection of U-LDM structures terminal nodes represent a contiguous textual span, we recombine the two parts of a non-contiguous segment in the BDU-tree (section 4) and DPT (section 5) using the concatenation relation (+). Concatenated nodes contain the complete f-structure information of the completed segment and are full-fledged BDUs, available for further "anaphoric anchoring".

All segments are returned to the discourse parser in an XML format, which includes the f-structure information as well as the textual spans for each sentential segment.

3 Discourse Parsing

The next step after segmentation is combining the segments into a BDU-tree according to the rules of discourse parsing, resulting in a discourse tree representing the sentence. After that, the BDU-tree is

[10] F-structures with SUBJ (subject) attributes are considered possible discourse segments since they typically encode an eventuality. However, because they do not introduce independent anchor points for future attachment, the f-structures corresponding to modal and auxiliary verbs are excluded.

combined with the Discourse Parse Tree, again according to the discourse parsing rules.

In discourse parsing, units of discourse are attached to an emergent discourse tree. Attachment always takes place on the right edge of the tree. The parser has to make two decisions: what unit to attach to and what relationship obtains between the incoming unit and the selected attachment site. The types of evidence are used to determine this include:

- **syntactic information**
  - subordinate/complement relations, parallel syntactic structure, topic/focus and centering information, syntactic status of re-used lexemes, preposed adverbial constituents, etc.
- **lexical information**
  - Information from lexical ontology: re-use of same lexeme, synonym/antonym, hypernym, participation in the same lexical frame as well as specific discourse connectives temporal and adverbial connectives indicating set or sub-set membership of any type *for example, specifically, alternatively*[11].
  - Modal information: realis status, modality, genericity, tense, aspect, point of view[12]),
- **Structural information** of both the local attachment point and of the BDU-tree
- **Presence of constituents of incomplete n-ary constructions on the right edge**
  - questions, initial greetings, genre-internal units such as sections and sub-sections, etc.

The combined weight of the evidence determines the attachment point and the attachment relation. Interestingly, the weight given to each type of information is different for attachment site selection and relationship determination. Lexical ontological information, for example, is generally more important in determining site, while semantic, syntactic and lexical "cue" information is more relevant in determining relationship.

In Polanyi et al. 2004, a small set of robust rules is given for determining the attachment site and relationship of an incoming BDU-tree to the existing parse tree of the discourse. In the present paper, which has as its focus understanding the relationship of sentential syntax to discourse structure, we concentrate on describing some fundamental aspects of the relationship between the sentential syntactic structure of an incoming sentence and its corresponding sentential discourse structure.

3.1 The Sentential Syntax - Discourse Structure Interface

In discourse parsing with LIDAS, the output of the XLE parser supplies functional constraints and roles for syntactic structures. The syntactic structure of a sentence accounts for the discourse-relevant relationships between segments within a sentence. LIDAS grammars exploit this information by mapping syntactic relations onto discourse relations. The LIDAS grammar formalism permits the parser to leverage the XLE's output by (1) checking positive or negative constraints (equality or inequality operators) (2) recursively searching f-structures for specified attributes (* and ? wildcards), (3) enforcing dependent constraints,[13] and (4) using Boolean connectives to combine constraints. (5) applying constraints universally or existentially to the set of matching f-structures. While LIDAS grammar rules can incorporate constraints that operate on all supported types of linguistic evidence, Table 3 gives three examples of rules that specify attachments based on the syntactic relationship between constituents.

1.	BDU-1,BDU-2: BDU-1/phi = BDU-2/ADJUNCT/link; → Right-Subordination.
2.	BDU-1,BDU-2: BDU-1/*/ADJUNCT/link = BDU-2/phi; → Subordination.
3.	BDU-1,BDU-2: BDU-1/*/{XCOMP|COMP|COMP-EX}/link = BDU-2/phi; → Context.

Table 3: Rules based on syntactic relationship.

Rule 1 captures one general case for preposed modifiers. Prepositional and adverbial phrases, often temporal modifiers, that precede the main clause they modify can either elaborate on the main clause or modify the context in which the main clause is interpreted. Lexical information is used to distinguish among different types of modifiers. Rule 2 expresses the general case of subordinate adjunct clauses and shows that the syntax for

[11] These expressions are used as input to specific lexically driven rules that indicate language or genre specific binary relations between BDUs as suggested by Webber, Joshi and colleagues in their work on D-LTAGs. (Webber and Joshi, 1998; Webber, Knott and Joshi, 1999; Webber, Knott, Stone and Joshi 1999.)

[12] See discussion Wiebe, 1994.

[13] For example, the following constraints show a dependency between (1) and (3),

(1) BDU-1/**(*)**/ADJUNCT/link = BDU-2/phi;
(2) BDU-2/ADJUNCT-TYPE = "relative";
(3) **$1**/SPEC/DET/DET-TYPE = "indef";

One or more sub-f-structures that match the wildcard in the first constraint are tested in the third constraint. This constraint is part of a rule of sentential discourse syntax used to identify non-restrictive relative clauses.

LIDAS' discourse rules allows for recursive search over f-structures by seeking adjunct phrases matching the incoming BDU anywhere within the f-structure of the attachment point. Rule 3, which shows a compact syntax for disjunctive constraint, builds a sentence level discourse relation, the Context Binary, that forms a complex context unit from its child constituents. Context Binaries are the general case for clausal complementizers.

In the next section, we discuss one of the basic discourse parsing rules.

3.2 Discourse Subordination

We will assume the following extended hierarchy of grammatical functions:[14] PRED > SUBJ > OBJ > COMP > ADJUNCT > SPEC.[15] Given this hierarchy we propose as a general principle of discourse construction that **promotion in the hierarchy means demotion in the discourse**. For example, if the SUBJ of the incoming unit of discourse refers to the same entity[16] as OBJ at the attachment node, then the relationship between them will be a subordination. In general, if an expression with grammatical function GF in a BDU ***B*** refers to the same (or a subcase of the same) entity as an expression with grammatical function GF' in an accessible antecedent BDU ***A***, where GF > GF', then ***B*** will be attached as a subordinate structure to ***A*** on the DPT. This principle is expressed as a rule in the grammar that fires if it is not superseded by other rules. For example, the Narrative rule, which coordinates event clauses, takes precedence over the Discourse Demotion Rule.

If the grammatical function hierarchy rule does not apply, but the BDU refers to a subclass[17] of the antecedent BDU, there is evidence for a subordination relation. For example, if the subject of the BDU stands in a part-of relationship with the subject of the antecedent BDU, we can conclude that the relationship is a subordination.

Table 4 gives the interaction of this rule with the hierarchy rule and the resulting relationships: G denotes whether a shift in the grammatical function hierarchy occurred, and L whether the shifted element refers to the same entity, part of that entity or to an entity that is larger. If more than one such expression can be found, we consider the expression in the incoming unit with the highest grammatical function.

	G^+	G^0	G^-
L^+	*S/C*[18]	*C*	*S*
L^0	*S*	*C*	*N/A*
L^-	*S*	*S*	*N/A*

Table 4. Interactions of the hierarchy and subcase rules.

The table is read as follows: take the expressions in the incoming unit that have a relationship with an expression in the attachment point. Let ***e*** be the expression among these that has the highest grammatical function, and ***f*** be the corresponding expression in the attachment point. If the grammatical function of ***e*** is higher than that of ***f***, we write G^+, if it is the same G^0, if it is lower G^-. Similarly, if ***e*** is a supercase of ***f***, we write L^+, if ***e*** and ***f*** refer to the same entity L^0, and if ***e*** is a subcase of ***f***, L^-.

For example

1. *U1: The man was wearing a coat.*
   U2: The Burberry looked brand new.

In this case we notice that the words *coat* and *Burberry* are such that $L<coat, Burberry> = L^-$ and we also notice that *Burberry* gets promoted to the subject of the incoming unit, while *coat* was the direct object of U1. Therefore $G<coat, Burberry> = G^+$. From Table 4 we correctly identify that U2 does indeed subordinate on U1.

Both L^- and G^+ give evidence for discourse subordination. Grammatical function demotion G^- is a less clear case.[19] Some mixed combinations suggest discourse coordination (as for the preservation case $<L^0, G^0>$) while others contribute too little

[14] PRED denotes the tensed verb. It plays a role in the following discussion because verbs can be nominalized.

[15] For the purposes of this hierarchy, grammatical function COMP includes the features COMP-EX and XCOMP. An element inside an expression with a grammatical function GF is itself in that position with respect to the elements that are not in that expression, although a separate ordering might exist between elements within the same expression. C.f. Obliqueness command in HPSG (Pollard and Sag, 1994).

[16] In LIDAS, no reference resolution is done. Identity of reference is approximated using lexical semantics.

[17] The notion of *subcase* as used here covers a number of different notions. An expression ***e*** is a subcase of ***f*** if (1) ***e*** is a set that is a subset of ***f***, (2) ***e*** is a subtype of ***f***, or (3) ***e*** is a part of ***f***, among other relations.

[18] Semantics would help to disambiguate this case. If the antecedent ***f*** is more specific than the anaphoric element ***e***, two cases are possible. Either ***e*** is used as a definite description referring to the same entity as ***f***, in which case the relation is a coordination, or ***e*** is used to denote a larger class of entities than ***f***, in which therelation is likely to be a subordination.

[19] As is understanding precisely what is involved from a discourse relation point of view with complex, mixed promotion/demotion phenomena (from the subject of an XCOMP to the adjunct of an OBJ, for example)

significant information to independently determine the discourse attachment (N/A cases). In those cases, evidence from other rules in the grammar determines the result.

4 BDU-Trees

Identifying the relationship of a BDU to the discourse in which it occurs is a complex parsing process involving lexical, semantic, structural and syntactic information. The first step in understanding the relationship of a given BDU to the extra-sentential discourse context is to understand the role a BDU plays within the sentence in which it occurs. As a first step of constructing a discourse parse-tree of the sentence, the XLE parse and sentential discourse rules are used to identify relationships between the BDUs within the sentence resulting in one or more *BDU-trees*. These small sentence-level discourse trees consist of one main clause and its subordinated clauses or preposed modifiers.[20]

The root node of a BDU-tree represents the non-subordinated material (often referred to as *active* material) within that BDU-tree, including at least the information of the main clause. The projection of the root node on the active leaves is referred to as the M-BDU (Main BDU). Only syntactic information from the M-BDU is used to attached the BDU-tree to the DPT. In general, a sentence yields as many BDU-trees as top-level coordinated clauses.

Consider, Example 2, a sentence taken from our corpus of Technical Reports:

2. *As a consequence, any hypertext link followed opened a new browser window, which we think of as a "Rabbit Hole" because the new window indicates to users that they are no longer navigating inside the slideshow, but are instead navigating the Web.*

The U-LDM segmentation of this sentence:

As a consequence | any hypertext link | followed | opened a new browser window | which we think of as a "Rabbit Hole" | because | the new window indicates to users | that they | are no longer navigating inside the slideshow | but | are instead navigating the Web.

[20] BDU-trees are very similar to the D-LTAG discourse segments (D-LDSs). However, BDU-trees include subordinated material, so they are typically larger than D-LDSs. Furthermore, because U-LDM sentence and discourse grammars are different, two BDU-trees may stand in a coordinated relationship in sentence grammar and be subordinated on the discourse level (cf. example 1).

For compound sentences, members of the top-level coordinate structures are attached independently, reflecting the fact that top-level coordinated clauses can escape the boundaries of the sentence when attaching to the main discourse structure. For example, the discourse parse of the following passage, illustrated in Example 3, consists of two sentences, five segments, four BDUs and three BDU-Trees that eventually form one DPT.

3. *S1: B1: The man soaked himself in the water.*
   S2: B2: It was warm and soothing B3 and he decided to linger a little longer than usual.

Despite the apparent syntactic coordination between the two main clauses, the two BDU-Trees in S2 show the independence of BDU-Tree/DPT attachments. B1 describes a punctual event on a main story line. B3 describes the next event. The semantic and aspectual information of the verb to soak in BDU-1 and of the copula in BDU-2 communicates a switch from an event-line to an embedded elaboration. The syntactic promotion of water from object of the preposition in the adjunct phrase *in the water* to the subject (through anaphoric reference) in the next segment indicates discourse subordination. Given L^0<water, it> and G^+<ADJUNCT/OBJ/PRED, SUBJ/PRED>, we find from table X, <L^0 ,G^+> → S. As a consequence, B2 is subordinated to B1. In the DPT, B3 is coordinated with B1 at a point above B2, despite being syntactically coordinated to it at the sentential level.

5 Global Discourse Construction

BDU-trees are attached to the DPT of the entire discourse as single units by computing the relationship between their M-BDUs and the accessible nodes aligned along the right edge of the DPT. Rules of discourse attachment that include those discussed for BDU construction as well as other genre and structural level rules are used in global DPT construction.

The parsing process at the Discourse Parse Tree (DPT) level works as follows. Once BDU-Trees have been constructed and are ready to be attached to the DPT, each node along the right edge is examined, and, through a set of discourse rules, an ordered set of active Discourse Constituent Units (DCUs) is produced, representing possible attachment points.[21] This set can then be pruned of its n

[21] Lexical information is the main source of evidence in attachment site determination while other sources contribute to a lesser degree. The opposite is true for determining attachment relations

lowest scoring constituents, according to a preset threshold or other criteria.

Example
Our group is developing new techniques for helping manage information for enhanced collaboration. We explore solutions for seamlessly connecting people to their personal and shared resources. Our solutions include services for contextual and proactive information access, personalized and collaborative office appliances, collaborative annotation, and symbolic, statistical, and hybrid processing of natural language. *Our team includes researchers with diverse backgrounds including: ubiquitous computing, computer-supported collaboration, HCI, IR , and NLP.*
Segmented Text
1. *Our group is developing new techniques for helping* **2.** *manage information for enhanced collaboration.* **3.** *We explore solutions for seamlessly connecting people to their personal and shared resources.* **4.** *Our solutions include services for contextual and proactive information access, personalized and collaborative office appliances, collaborative annotation, and symbolic, statistical, and hybrid processing of natural language.* **5.** *Our team includes researchers with diverse backgrounds* **6.** *including:* **7.** *ubiquitous computing, computer-supported collaboration, HCI, IR , and NLP.*
Rules
1-2 Intrasentential XCOMPS --> CX CX(1,2)-3 Demotion 2 [Pres. Progressive to Simple Present] [Our group --> We] [Same Verb Class] --> S 3-4 Promotion [Solutions from OBJ to SUBJ] --> S 5-6 Relative Adjunct with Null Determiner --> S 6-7 Colon + OBJ linked to NP segment --> CX CX(1,2)-5 Synonym Subjects, Same Tense --> C[22]

Table 5. Analyzed Webpage Example

In the second stage, attachment rules are checked against possible attachment sites. Rules that fire successfully attach the BDU-Tree to the DPT at the chosen site with the relationship specified by the rule. Local semantic, lexical and syntactic information is then percolated into the parent DCU. If multiple attachments at different sites are possible, ambiguous parses are generated; less preferred attachments are discarded and the remaining attachment choices generate valid parse trees.

Once a BDU-tree has been attached, its leaves become terminal nodes of the DPT and nodes on its right edge are accessible for continuation and may serve as contextual anchors for subsequent sentences. Table 5 shows an example text taken from the FXPAL Webpage describing our research group, a segmentation of the text and the DPT constructed following the rules. Figure 1 gives a screenshot of the resulting tree.

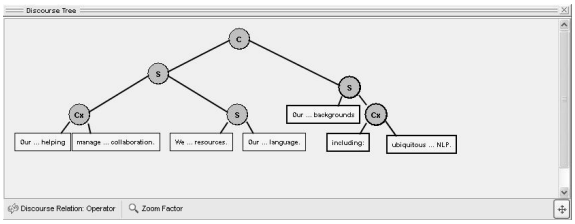

Figure 1: Tree of Analyzed Webpage Example

6 Comparison with D-LTAG

LIDAS is a purely symbolic discourse parser most similar to the D-LTAG parser described in (Forbes et. al. 2003). The overall structures of the LIDAS and D-LTAG parsers are almost identical. There are a number of apparently significant differences between the underlying theories although some of these may turn out to be notational variants after more extensive analysis.

An important difference between D-LTAG and U-LDM derives from the fundamental question of segmentation; D-LTAG segments are larger than our segments, corresponding more or less to BDU-trees (but cf. footnote 20). In D-LTAG, because only one grammar formalism governs both discourse and sentence level parsing, continuation can also take place on parts of segments as defined by sentence-level syntactic relations. Under the U-LDM, which employs independent sentential and discourse grammars, only segments are potential anchors for continuation. Not only because we use an external syntactic parser as an oracle that informs segment attachment on the BDU-tree, but more importantly because sentential syntax can be overridden by discourse syntax in some cases.

Another basic difference between the two approaches is that D-LTAG builds its initial and auxiliary trees around connectives. Every discourse relation is expressed by a, possibly empty, connective. In U-LDM, connectives give evidence about the possible discourse relations, as do other parts of the sentence, but they do not solely determine discourse relation. We can thus account correctly

[22] Discourse parsing is not unambiguous and different analyses may apply. So, here the same subject, change in progressive feature, and different verb class, which also apply would create an S. Future research is needed to understand precisely which rules take precedence.

for cases in which the wrong connective is selected to express a semantic relationship among segments.

Lastly, our parser is meant to be incremental at the discourse level, whereas the D-LTAG parser appears to operate on the discourse as a whole. [23]

7 Conclusion

In this paper we described a novel approach to discourse segmentation and discourse structure analysis that unifies sentential syntax with discourse structure and argued that most of the information needed to assign a structural description to a text is available from sentential syntactic parsing, sentential semantic analysis and relationships among words captured in a lexical ontology such as WordNet. The U-LDM discourse rules and parsing strategies presented here are a first step. We have tested out these rules in analyzing a corpus of over 300 Technical Reports that have been summarized under the PALSUMM System that operates on top of LIDAS. (Polanyi et al 2004; Thione et al 2004) Much work remains to be done. Understanding the complex inter-relationships between rules is a formidable task. Critically important, too, is to unify the semantically motivated structural analyses presented here with an explicit S-DRT type formal semantic account of discourse semantics. However, we believe that the results presented here represent an important advance in understanding the nature of natural language texts.

8 References

Nicholas Asher. 1993. Reference to Abstract Objects in English: A Philosophical Semantics for Natural Language Metaphysics. Kluwer Academic Publishers.

Nicholas Asher and Alex Lascarides. 2003. Logics of Conversation. Cambridge University Press.

Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind Joshi and Bonnie Webber. 2003. D-LTAG System - Discourse Parsing with a Lexicalized Tree-Adjoining Grammar, Journal of Language, Logic and Information, 12(3).

Barbara Grosz and Candace Sidner. 1986. Attention, Intention and the Structure of Discourse. Computational Linguistics 12:175-204.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. Text 8(3)243-281.

[23] In the current LIDAS implementation, we do not represent ambiguity directly, but implement a greedy parsing algorithm with backtracking. The non-locality of the D-LTAG parser as described in Forbes et. al. 2003 may likewise be a consequence of their current implementation.

Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. In Advances in Automatic Text Summarization. I. Mani and Mark Maybury (eds), 123-136, The MIT Press.

Daniel Marcu. 2000. The Theory and Practice of Discourse Parsing and Summarization. The MIT Press. Cambridge, MA.

John Maxwell and Ronald M. Kaplan. 1989. An overview of disjunctive constraint satisfaction. In Proceedings of the International Workshop on Parsing Technologies, Pittsburgh, PA.

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

Livia Polanyi. 2004. A Rule-based Approach to Discourse Parsing. In Proceedings of SIGDIAL '04. Boston MA.

Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, PA.

Stefan Riezler, Tracy H. King, Richard Crouch, and Annie Zaenen. 2003. Statistical Sentence Condensation using Ambiguity Packing and Stochastic Disambiguation Methods for Lexical-Functional Grammar. In Proceedings of HLT-NAACL'03, Edmonton, Canada.

Radu Soricut and Daniel Marcu. 2003. Sentence Level Discourse Parsing using Syntactic and Lexical Information. In Proceedings of HLT/NAACL'03, May 27-June 1, Edmonton, Canada

Thione, Gian Lorenzo, Martin van den Berg, Chris Culy and Livia Polanyi. 2004a. Hybrid Text Summarization: Combining external relevance measures with Structural Analysis. Proceedings ACL Workshop Text Summarization Branches Out. Barcelona.

Thione, Gian Lorenzo, Martin van den Berg, Chris Culy and Livia Polanyi. 2004b. LiveTree: An Integrated Workbench for Discourse Processing. Proceedings ACL Workshop on Discourse Annotation. Barcelona.

Bonnie Webber and Aravind Joshi, 1998. Anchoring a lexicalized tree-adjoining grammar for discourse. COLING/ACL Workshop in Discourse Realtions and Discourse Markers. Montreal, Canada. 86-92.

Bonnie Webber, Alistair Knott and Aravind Joshi. 1999a. Multiple discourse connectives in a lexicalized grammar for discourse. In 3rd Int'l Workshop on Computational Semantics. Tilburg. 309-325.

Bonnie Webber, Alistair Knott, Matthew Stone and Aravind Joshi. 1999b. Discourse Relations: A Structural and Presuppositional Account Using Lexicalized TAGS. 37th ACL. College Park, MD. 41-48.

Wiebe, Janyce M. 1994. Tracking point of view in narrative. *Computational Linguistics 20* (2): 233-287.