DISCOURSE STRUCTURE: SWINGS AND ROUNDABOUTS

BONNIE WEBBER
University of Edinburgh

RASHMI PRASAD
University of Pennsylvania

ABSTRACT

The goal of understanding how discourse is more than a sequence of sentences has engaged researchers for many years. Researchers in the 1970's attempted to gain such understanding by identifying and classifying the phenomena involved in discourse. This was followed by attempts in the 1980s and early 1990s to explain discourse phenomena in terms of theories of abstract structure. Recent efforts to develop large-scale annotated discourse corpora, along with more lexically grounded theories of discourse are now beginning to reveal interesting patterns and show where and how early theories might be revised to better account for discourse data.

In the sciences, theory and data often compete for the hearts and minds of researchers. In linguistics, this has been as true of research on discourse structure as of research on syntax. Researchers' changing engagement with data versus theory in discourse structure – and the hope for more progress by engaging with both of them – is the subject of this brief paper.

[1] THE 70S: A FOCUS ON DATA

The 1970's saw a focus on data, with *Cohesion in English* (Halliday and Hasan 1976) an important milestone. This volume catalogued linguistic features in English that impart *cohesion* to a text, by which Halliday and Hasan meant the network of lexical, referential, and conjunctive relations which link together its different parts. These relations contribute to creating a text from disparate sentences by requiring words and expressions in one sentence be interpreted by reference to words and expressions in the surrounding sentences and paragraphs.

Of particular interest here are *conjunctive elements*, which Halliday and Hasan took to signal how an upcoming sentence is related to what has been said before. Such conjunctive elements include both *co-ordinating and sub-ordinating conjunctions* and what they call *conjunctive adjuncts* (eg, adverbs such as *but, so, next, accordingly, actually, instead, besides, etc.*; and prepositional phrases such as *as a result, in addition, in spite of that, in that case, etc.*).

More specifically, a conjunctive element is taken to convey a cohesive relation between its matrix sentence and that part of the surrounding discourse that supports its effective decoding – resolving its reference, identifying its sense, or recovering missing material needed for its interpretation. Three examples of conjunctive elements can be found in this extract given from *Meeting Wilfred Pickles*, by Frank Haley.

- (1)
 - a. Then we moved into the country, to a lovely little village called Warley.
 - b. It is about three miles from Halifax.
 - c. There are quite a few about.
 - d. There is a Warley in Worcester and one in Essex.
 - e. But the one not far out of Halifax had had a maypole, and a fountain.
 - f. By this time the maypole has gone, but the pub is still there called the Maypole.

Halliday and Hasan labelled the adverb *then* in (1-a) a SIMPLE SEQUENTIAL TEMPORAL CONJUNCTION (like *next*), which is a type of SEQUENTIAL TEMPORAL CONJUNCTION, which is itself a type of TEMPORAL CONJUNCTION. It is decoded as conveying a simple sequential temporal relation to something in the preceding text (not provided here). *But* in (1-e) was labelled a SIMPLE CONTRASTIVE ADVERSATIVE CONJUNCTION (like *and*), which is a type of a CONTRASTIVE ADVERSATIVE CONJUNCTION, itself a type of ADVERSATIVE CONJUNCTION. *But* is decoded here as conveying a simple contrastive adversative relation to (1-d). The final conjunctive element in this text, *by this time*, Halliday and Hasan labelled a COMPLEX TERMINAL TEMPORAL CONJUNCTION (like *until then*), which is a type of TERMINAL TEMPORAL CONJUNCTION, which is another type of TEMPORAL CONJUNCTION. *By this time* is decoded as conveying a complex temporal relation to (1-a).

Although Halliday and Hasan provided a very elaborate taxonomy of conjunctive elements in terms of the hierarchy of detailed labels for conjunctive relations illustrated above, they did not embed it in any kind of theoretical framework that would explain, for example, how meaning is projected systematically from a conjunctive relation with a given label or what “surrounding text” a link can be made to. Without such a theoretical framework, it was difficult to make use of their data analysis in the systematic way required for computational applications.

[2] THE 80S AND 90S: A FOCUS ON THEORY

Providing a theoretic framework for what can be termed *lexically grounded* discourse relations was not, however, what researchers in the 1980s and 1990s were concerned with. Rather, they aimed to provide a complete theoretical account of a text in terms of *abstract* discourse relations. Such accounts included *Rhetorical Structure Theory* (hereafter, RST) developed by Mann and Thompson (1988); a theory (hereafter, GS) developed by Grosz and Sidner (1986) that posited three separate but isomorphic discourse structures – an intentional structure, a linguistic structure and an attentional structure; the *Linguistic Discourse Model* (hereafter, LDM) developed by Polanyi and her colleagues (Polanyi 1988; Polanyi and van den Berg 1996); *Relational Discourse Analysis* (hereafter RDA) developed by Moser and Moore (1996) as a way of reconciling RST and GS; and *Structured Discourse Representation Theory* (hereafter SDRT) developed by Asher and Lascarides (2003) as an extension to DRT (Kamp and Reyle 1993) to account for how discourse relations arise from what the authors call *commonsense entailment* (Lascarides and Asher 1993). The abstract discourse relations used in these theories included both semantic relations between the facts, beliefs, situations, eventualities, etc. described in a text (more often called *informational relations*) and pragmatic relations between what a speaker is trying to accomplish with one part of a text with respect to another (more often called *intentional relations*). Together, these are often simply called *discourse relations*.

Unlike the lexically-grounded discourse relations of Halliday and Hasan (1976), theories of abstract discourse relations see comprehensive structure (based on discourse relations) underlying a text, just as theories of syntax see a comprehensive syntactic structure underlying a sentence. In particular, theories of abstract discourse relations assume, as in formal grammar, a set of terminal elements (called either *elementary* or *basic* discourse units). A discourse relation holding between adjacent elements joins them recursively into a larger unit, with a text being recursively analysable down to its terminal elements. Such an analysis covers the entire text, just as in syntax a single parse tree or dependency analysis covers an entire sentence. Also as in syntax, the analysis is essentially a tree structure since, for the most part, it is also assumed that no discourse element is part of more than one larger element. Where these theories differ is in (1) the specific types of relations they take to hold between units; (2) the amount of attention they give to, eg, how a hearer establishes the relation that holds between two units; (3) whether they take there to be separate but related *informational* and *intentional* discourse structures (GS, RDA) or a single structure (RST, LDM, SDRT); and (4) how to provide a compositional syntactic-semantic interface that would systematically interpret a discourse unit in terms of the interpretations assigned to its component parts.

These theories were not without practical application, and came to under-

pin work in Natural Language Generation (Marcu 1996; Mellish et al. 1998; Moore 1995) and document summarization (Bosma 2004; Marcu 1998, 2000). However, these theories were based on very little data, and problems began to be noticed early on. For example, Moore and Pollack (1992) noticed that the same piece of text could simultaneously be given different RST analyses, each with a different structure. These were not alternative analyses that could be disambiguated: all of them seemed simultaneously appropriate. But if this was the case, what was the consequence for associating a text with a single discourse structure (or even structurally isomorphic intentional and informational structures)?

Elsewhere, Scott and de Souza (1990) and later Carlson et al. (2003) pointed out that the meaning conveyed by a sequence of sentences or a complex sentence could also be conveyed by a single clause. But if this was the case, did it make sense to posit an independent existence for elementary discourse units?

And questions about discourse having an underlying recursive tree-like structure were raised by Wiebe (1993) based on examples like

- (2) a. The car was finally coming toward him.
- b. He [Chee] finished his diagnostic tests,
- c. feeling relief.
- d. *But then* the car started to turn right.

The problem she noted was that the discourse connectives *but* and *then* appear to link clause (2-d) to two different things: *then* to clause (2-b) in a SEQUENCE relation – i.e., the car starting to turn right being the next relevant event after Chee’s finishing his tests – and *but* to a grouping of clauses (2-a) and (2-c) – i.e., reporting a contrast between, on the one hand, Chee’s attitude towards the car coming towards him and his feeling of relief and, on the other hand, his seeing the car turning right. But a structure with one subtree over the non-adjacent units (2-b) and (2-d) and another over the units (2-a), (2-c) and (2-d) again is not itself a tree. Did such examples really raise a problem for postulating an underlying comprehensive tree-like structure for discourse, and if they did, what kind of comprehensive structure, if any, did discourse have?

[3] THE NEW CENTURY: RE-FOCUSSING ON DATA

While the problems noted above weren’t enough to refute theories of abstract discourse relations, two distinct currents gaining momentum in sentence-level syntax through the 1990s have led discourse research to re-focus on data again in the new century. These were the emergence of (1) part-of-speech and syntactically annotated corpora such as the Penn TreeBank (Marcus et al. 1993); and (2) lexicalized grammars, in which syntactic contexts were associated with words directly rather than indirectly through phrase structure rules (eg, syntactic con-

texts in the form of tree fragments in Lexicalized Tree-Adjoining Grammar (LTAG) (Schabes 1990; XTAG-Group 2001) and in the form of complex categories in Combinatory Categorical Grammar (CCG) (Steedman 1996, 2000)).

In Section 3.1, we will make some general remarks about annotated corpora, followed by a brief description in Section 3.2 of a lexicalized grammar for discourse, and then finally, in Section 3.3, a bit about a particular annotated discourse corpus – the Penn Discourse TreeBank – that was stimulated by this work to return to a focus on lexically-grounded discourse relations. We will close with some predictions about the future.

[3.1] *Annotated Discourse Corpora*

While the automatically generated, manually corrected part-of-speech and syntactic annotation that makes up the Penn TreeBank (PTB) Wall Street Journal Corpus was developed as a community-accepted “gold standard” on which parsers and parsing techniques could be *evaluated* (Marcus et al. 1993), the PTB soon became a basis for *inducing* parsers using statistical and machine learning techniques that were much more successful in wide-coverage parsing than any ones previously developed. Similar techniques based on appropriately annotated corpora enabled the development of other language technology, including part-of-speech taggers, reference resolution procedures, semantic role labellers, etc., which are beginning to be used to improve performance of applications in information retrieval, automated question answering, statistical machine translation, etc.

Eventually, researchers turned to consider whether annotated discourse corpora can yield similar benefits by supporting the development of technology that required sensitivity to discourse structure, such as in extractive summarization, where one identifies and includes in a summary of one or more source texts, only their most important sentences, along with, perhaps, other sentences needed to make sense of them. This was one of the motivations behind the creation of the RST Corpus (Carlson et al. 2003), comprising 385 documents from the Penn TreeBank corpus that have been manually segmented into elementary discourse units, linked into a hierarchy of larger and larger units, and annotated with RST relations taken to hold between linked units. It was also the acknowledged reason for developing an LDM-annotated corpus (Polanyi et al. 2004). Another corpus was annotated according to RDA (Moser and Moore 1996), with an aim of improving the quality of Natural Language Generation (NLG) – in particular, to identify which of several syntactic variants is most natural in a given context – information that can then be incorporated into the sentence planning phase of NLG (Di Eugenio et al. 1997). (Prasad et al. (2005) discusses how the Penn Discourse TreeBank, to be discussed in Section 3.3, can also be used for this purpose.)

While the promise of benefits for Language Technology have helped attract funding for annotated discourse corpora, such corpora serve other objectives as

well. For example, researchers want to use them to test hypotheses about the effect (or co-dependency) of discourse structure on other aspects of language such as argumentation (Stede 2004; Stede et al. 2007) or reference resolution (cf. work done at the University of Texas on a corpus annotated according to SDRT (Stede et al. 2007)).

But a fundamental reason for developing annotated discourse corpora is to enable us to advance towards a theoretically well-founded understanding of discourse relations – ie, of how they arise from text, or of constraints on their possible arguments, or of how the resulting structures pattern – that is well-grounded in empirical data. This was a main goal behind the development of the Discourse GraphBank (Wolf and Gibson 2005).¹ And already there has been a real advance in discourse theory based on empirical data (Stede 2008): Problems with annotating RST relations in the Potsdam Commentary Corpus (Stede 2004; Stede et al. 2007) led Stede to examine in detail the notion of *nuclearity* that has been fundamental to RST as a theory but is problematic in practice. The result is an argument for nuclearity as primitive only for *intentional discourse relations*. For *informational relations*, nuclearity is best discarded in favor of independently motivated notions of *discourse salience* or *prominence* (associated with entities) and *discourse topic*. Before turning to a brief description of the Penn Discourse TreeBank (Miltsakaki et al. 2004; Prasad et al. 2004; Webber 2005) and some things we have already learned from the process of creating it, we will briefly describe the lexicalized approach to discourse that led to the creation of this lexically-grounded corpus.

[3.2] *Discourse Lexicalized Tree-Adjoining Grammar (D-LTAG)*

D-LTAG is a lexicalized approach to discourse relations, which aims to provide an account of how lexical elements (including phrases) anchor discourse relations and how other parts of the text provide arguments for those relations (Webber et al. 2003; Webber 2004). D-LTAG arose from a belief that language has only a limited number of ways to convey relations between things. For example, within a clause,

- a verb or preposition can convey a relation holding between its arguments (eg, the cat *ate* the cheese; the cat *in* the hat), as can some nouns and adjectives (eg, the *love* of man for his family; an *easy* mountain to climb);
- adjacency of two or more elements can convey implicit relations holding between them, as in *soup pot*, *soup pot cover*, *aluminum soup pot cover adjustment screw*, etc;
- an anaphoric expression conveys a relation between all or part of its denotation and some element of the surrounding discourse – an *IDENTITY* rela-

[1] (Webber 2006) argues against the strong claims (Wolf and Gibson 2005) make about discourse structure based on this annotation, and many of them have subsequently been withdrawn (Kraemer and Gibson 2007).

tion in the case of coreference, other relations in the case of comparative anaphora such as *the smaller boys*;

- intonation and information structure can convey relations of CONTRAST between discourse elements and/or their different roles with respect to information structure (Steedman 2007).

If one assumes that many of the same means are operative outside the clause as within it, then it makes sense to adopt a similar approach to discourse analysis as to syntactic analysis. Since lexicalized grammars seemed to provide a clearer, more direct handle on relations and their arguments at the clause-level, D-LTAG adopted a lexicalized approach to discourse based on lexicalized Tree Adjoining Grammar (Schabes 1990). A lexicalized TAG (LTAG) differs from a basic TAG in taking each lexical entry to be associated with the set of elementary tree structures that specify its local syntactic configurations. These structures can be combined via either *substitution* or *adjoining*, to produce a complete sentential analysis.

The elementary trees of D-LTAG are anchored by discourse connectives whose substitution sites correspond to their arguments. These can be filled by anything interpretable as an *abstract object* (ie, as a proposition, fact, eventuality, situation, etc.). Elements so interpretable include discourse segments, sentences, clauses, nominalisations and demonstrative pronouns. Adjacency in D-LTAG is handled by an elementary tree anchored by an *empty connective*.

As with a sentence-level LTAG, there are two types of elementary trees in D-LTAG: *Initial trees*, anchored by *structural connectives* such as subordinating conjunctions and subordinators (eg, *in order to*, *so that*, etc.) illustrated by the trees labeled α :so and α :because_mid in Figure 1(a), and *auxiliary trees*, anchored by a coordinating conjunction or an *empty connective* or by a discourse adverbial, illustrated by the trees labeled β :but and β :then in Figure 1(a).

Both the initial trees of structural connectives and the auxiliary trees of coordinating conjunctions and the *empty connective* reflect the fact that both their arguments are provided through structure – in the case of initial trees, the two substitution arguments labelled with $\downarrow$ in α :so and α :because_mid, and in the case of auxiliary trees, the one substitution argument labelled with $\downarrow$ and the adjunction argument labelled with $*$, as in β :but. In contrast, a discourse adverbial is *anaphoric* – with the discourse relation, part of its semantics, but with one of its arguments coming from the discourse context by a process of anaphor resolution. Its other argument is provided structurally, in the form of its matrix clause or sentence. That discourse adverbials such as *instead*, *afterwards*, *as a result*, etc. are *anaphoric*, differing from structural connectives in getting their second argument from the discourse context, is argued on theoretical grounds in (Webber et al. 2003) and on empirical grounds in (Creswell et al. 2004). It also echoes in part the claim of Halliday and Hasan (1976), noted in Section 1, that all conjunctive el-

(a) Derived Tree for Example (3).

(b) Derivation Tree for Example (3)

FIGURE 1: Tree analyses of Example (3)

ements were interpreted in this way. Justification of the anaphoric character of discourse adverbials is given in (Forbes 2003) and (Forbes-Riley et al. 2006).

Discourse relations arising from both structural and anaphoric connectives can be seen in the D-LTAG analysis of Example (3) below.

- (3) John loves Barolo.
 So he ordered three cases of the '97.
 BUT he had to cancel the order
 BECAUSE he THEN discovered he was broke.

Figure 1(a) illustrates both the starting point of the analysis — a set of elementary trees for the connectives (*so*, *but*, *because*, *then*) and a set of leaves (*T1-T4*) for the four clauses in Example (3) without the connectives — and its ending point, the *derived tree* that results from *substituting* at the nodes in α :*because_mid*, α :*so* and β :*but* marked $\downarrow$ and *adjoining* the trees β :*but* and β :*then* at their nodes marked *. These operations of *substitution* and *adjoining* are shown as solid and dashed lines respectively in Figure 1(b), in what is called a *derivation tree*. The numbers on the arcs of the derivation tree refer to the node of the tree that an operation has been performed on. For example, the label 1 on the solid line from α :*so* to *T1* means that *T1* has substituted at the leftmost node of the tree α :*so*. The label 3 refers to the node that is third from the left. The label 0 on the dashed line from α :*so* to β :*but* means that β :*but* has adjoined at the root of α :*so*.

The structural arguments of a connective can come about through either substitution or adjoining. The derived tree in Figure 1(a) shows the two structural arguments of *so*, *but* and *because* as their left and right sisters. The derivation tree in Figure 1(a) shows both arguments to *so* and *because* coming from substitution, with one structural argument to *but* coming from substitution and the other coming through adjoining. Finally, *then* only has one structural argument, shown in Figure 1(a) as its right sister. Figure 1(b) shows it coming from adjoining. The dotted line in Figure 1(a) shows *then* linked anaphorically to the clause that gives rise to its first argument. More detail on both the representation of connectives and D-LTAG derivations is given in (Webber et al. 2003). A preliminary parser producing such derivations is described in (Forbes et al. 2003) and (Webber 2004).

Compositional interpretation of the derivation tree produces the discourse relation interpretations associated with *because*, *so* and *but*, while anaphor resolution produces the second argument to the discourse relation interpretation associated with *then* (ie, the ordering event), just as it would if *then* were paraphrased as *soon after that*, with the pronoun *that* resolved anaphorically. Details on this syntactic-semantic interface are given in (Forbes-Riley et al. 2006).

Although D-LTAG produces only analyses in the form of trees, Webber et al. (2003) recognized that occasionally the same discourse unit participates in one relation with its left-adjacent material and another distinct relation with its right-

adjacent material, as in

- (4) AS SOON AS **the tremor passed**, *many people spontaneously arose and cheered*,
AS THOUGH **it had been a novel kind of pre-game show**.²
- (5) WHEN **Ms. Evans took her job**, *several important divisions that had reported
to her predecessor weren't included* PARTLY BECAUSE **she didn't wish to be a
full administrator**.

In Example (4), the main clause “many people spontaneously arose and cheered” serves as ARG1 to both the subordinating conjunction AS SOON AS on its left and the subordinating conjunction AS THOUGH on its right. Example (5) shows a similar pattern. Such examples, however, would require relaxing substitution constraints in D-LTAG that the same tree only substitutes into a single site. This has not yet been done. It should be clear that D-LTAG is only a conservative extension of the theories mentioned in Section 2 in that it shares their assumptions that a text can be divided into discourse units corresponding to clauses, with a discourse analysis covering those units – hence the work on developing a discourse parser for D-LTAG (Forbes et al. 2003; Webber 2004). The main way in which D-LTAG diverges from these other theories is in anchoring discourse relations in, on the one hand, structural connectives and adjacency, and on the other, anaphoric connectives. The latter provide additional relations between material that is not necessarily adjacent, but not in a way that changes the complexity of discourse structure: D-LTAG analyses are still trees.

[3.3] The Penn Discourse TreeBank

The Penn Discourse TreeBank (PDTB) annotates *discourse relations* in 2304 articles of the Wall Street Journal corpus (Marcus et al. 1993) in terms of *discourse connectives*, the *minimal text spans* that give rise to their arguments, and the attribution of both the connectives and their arguments (Dinesh et al. 2005; Prasad et al. 2007). For example, in Example (6)

- (6) *Factory orders and construction outlays were largely flat in December* WHILE
purchasing agents said **manufacturing shrank further in October**.

both ARG1 and the connective *while* are attributed to the writer, while ARG2 is attributed to someone else via the attributive phrase “purchasing agents said”, which is not included in that argument.

Primarily two types of connectives have been annotated in the PDTB: *explicit connectives* and *implicit connectives*, the latter being inserted between adjacent

[2] Both these examples come from the Penn Discourse TreeBank (Section 3.3) and reflect the actual annotation of these connectives. In all PDTB examples, ARG1 is in italics and ARG2 in boldface, and the connective is underlined.

CONNECTIVE	ADJACENT	NON-ADJACENT	TOTAL
S-medial <i>so</i>	147	2	149
S-medial <i>but</i>	213	0	213
Total	360	2	362
S-initial <i>So</i>	89	22	111
S-initial <i>But</i>	347	63	410
Total	436	85	521

FIGURE 2: S-medial vs. S-initial Connectives

paragraph-internal sentences not related by an explicit connective. As in D-LTAG (Section 3.3), explicit connectives include coordinating conjunctions, subordinating conjunctions and discourse adverbials. The argument associated syntactically with the discourse connective is conventionally referred to as ARG2 (eg, the subordinate clause of a subordinating conjunction) and the other argument as ARG1. Because annotators were asked to annotate only the *minimal span* associated with an argument, they were also allowed (but not required) to indicate spans adjacent to ARG1 and ARG2 that were relevant to them but still supplementary, using the tags SUP1 and SUP2.

A preliminary version of the PDTB containing the annotation of 18505 explicit connectives was released in April 2006 (PDTB-Group 2006), and received over 120 downloads. The completed PDTB (Version 2.0) was released by the Linguistic Data Consortium (LDC) in February 2008, and includes annotation of all implicit connectives as well, along with a hierarchical semantic annotation of both explicit and implicit connectives (Miltsakaki et al. 2008). More information on the PDTB can be found at its homepage (<http://www.seas.upenn.edu/~pdtb>).

Although annotation has sometimes been a challenge, it has nevertheless begun to yield some useful observations. One such observation is that the position of ARG1 can differ significantly, depending on whether its associated connective occurs sentence-medially (*S-medial*) or sentence-initially (*S-initial*). Figure 2 contrasts the instances of S-medial and S-initial *so* and *but* in the corpus³ with respect to whether their ARG1 spans the immediately preceding clause(s) or sentence(s) (**Adjacent**) or not (**Non-adjacent**). This difference in patterning between medial instances of *but* and *so* and sentence-initial instances, is statistically significant ($<.0001$).

The difference will certainly be relevant to those researchers interested in developing the technology to automatically recognize the arguments to discourse

[3] With *But*, figures are based on the first 213 of 1188 S-medial instances in the corpus and the first 410 of 2124 S-initial instances.

connectives (cf. [Wellner and Pustejovsky 2007](#)). But it is not just the distance of ARG1 from its connective: it is also a difference in function. As with S-initial *but* and *so*, ARG1 of S-initial discourse adverbials like *instead* can also be found at a distance – eg,

- (7) On a level site you can provide a cross pitch to the entire slab by *raising one side of the form*, but for a 20-foot-wide drive this results in an awkward 5-inch slant across the drive’s width. **INSTEAD, make the drive higher at the center.**
(*Reader’s Digest New Complete Do-it-yourself Manual*, p. 154)
- (8) *If government or private watchdogs insist, however, on introducing greater friction between the markets (limits on price moves, two-tiered execution, higher margin requirements, taxation, etc.), the end loser will be the markets themselves.* **INSTEAD, we ought to be inviting more liquidity with cheaper ways to trade and transfer capital among all participants.**

But of the 48 occurrences of S-initial *Instead* in the corpus, ARG1 spans text other than the main clause in 26 (54.2%), independent of its position with respect to the connective. In Example (8), ARG1 spans the subordinate clause, and in Example (9), the gerund complement of an appositive NP:

- (9) 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.**

This suggests that ARG1 of a discourse adverbial like *instead* comprises a span with a specific semantic character — here, one that can be interpreted as something for which an alternative exists ([Webber 2004](#)). The span can be anywhere in the sentence and serve any role. In contrast, ARG1 of S-initial *So* or *But* comprises a span at the same level of discourse embedding as ARG2, with any intervening material appearing to serve a supporting role to the span identified as ARG1, as in

- (10) *The 40-year-old Mr. Murakami is a publishing sensation in Japan. A more recent novel, “Norwegian Wood” (every Japanese under 40 seems to be fluent in Beatles lyrics), has sold more than four million copies since Kodan-sha published it in 1987.* **BUT he is just one of several youthful writers – Tokyo’s brat pack – who are dominating the best-seller charts in Japan.**

- (11) *It is difficult, if not impossible, for anyone who has not pored over the thousands of pages of court pleadings and transcripts to have a worthwhile opinion on the underlying merits of the controversy. Certainly I do not. So **we must look elsewhere for an explanation of the unusual power this case has exerted over the minds of many, not just in Washington but elsewhere in the country and even the world.***

Additional (albeit incomplete) evidence for this role of the intervening text comes from the frequency with which some or all of the intervening material has been annotated SUP1 in the PDTB, even though the annotators were not required to be systematic or rigorous in their use of this label.

This pragmatic/intentional sense of a supporting role (or a difference in *focality* or *discourse salience*) appears to be what Mann and Thompson (1988) had in mind with *presentational* rhetorical relations in which one argument (the *satellite*) supported the other argument (the *nucleus*) – and what Grosz and Sidner (1986) had in mind when they posited a *dominance* relation in which one discourse segment supported the discourse purpose of another. It also appears related to what Blühdorn (2007) and Ramm and Fabricius-Hansen (2005) refer to as *subordination* in discourse. It has not been directly annotated in the PDTB, which has focussed on the arguments to explicit and implicit connectives. On the other hand, we should be able to gather additional evidence related to this issue once the annotation of implicit connectives in the PDTB has been adjudicated, at which point we can survey what relations annotators have taken to hold between ARG1 and material intervening between it and its associated connective. Such empirical evidence of something like *intentional structure* would be very exciting to find.

A second observation relates to other linguistic devices that convey discourse relations, besides the discourse connectives annotated in the PDTB. In annotating implicit connectives, we noticed different cases whose paraphrase in terms of an explicit connective sounded redundant. A closer look revealed systematic non-lexical indicators of discourse relations, including:

- cases of S-initial PPs and adjuncts with anaphoric or deictic NPs such as *at the other end of the spectrum*, *adding to that speculation*, which convey a relation (eg, CONTRAST and ADDITIONAL INFORMATION in the above two cases) in which the immediately preceding sentence provides a referent for the anaphoric or deictic expression, which thereby makes it available as ARG1.
- cases of equative *be* in which the immediately preceding sentence provides a referent for an internal anaphoric argument of a relation-containing NP in subject position – eg, the CONSEQUENT relation conveyed by *the effect is* (ie, *the effect [of that] is*) in

- (12) *The New York court also upheld a state law, passed in 1986, extending for one year the statute of limitations on filing DES lawsuits. THE EFFECT IS THAT lawsuits that might have been barred because they were filed too late could proceed because of the one-year extension.*

Identification of these alternative ways of conveying discourse relations (labelled *ALTLEX* in the PDTB) will allow for a more complete annotation of discourse relations in other corpora, including corpora in languages other than English.⁴

A third observation relates to the distinction that [Blühdorn \(2007\)](#) has drawn between discourse hierarchy associated with the pragmatic/intentional concept of *focality* versus discourse hierarchy associated with the syntactic/semantic concept of *constituency*. It is clear that both some sort of *intentional structure* and some sort of *informational structure* are needed in discourse, just as was suggested in the theories of [Grosz and Sidner \(1986\)](#), [Moore and Pollack \(1992\)](#), and Moser and Moore (1996). It is just that the data seem to tell a somewhat different story of the properties of these structures than those assumed in these earlier theories of abstract discourse relations.

Our previous observation about the status of material intervening between ARG1 and S-initial *But* and *So* related to a discourse hierarchy associated with *focality*, with the intervening material playing a supplemental role in the discourse. Our observation here relates to *constituency* and is a consequence of the procedure used in annotating the PDTB. This differed from the procedure used in annotating the RST corpus ([Carlson et al. 2003](#)), the LDM corpus ([Polanyi et al. 2004](#)) and the GraphBank corpus ([Wolf and Gibson 2005](#)), where annotators first marked up the text with a sequence of *elementary discourse units* that exhaustively covered it (just as a sequence of words and punctuation exhaustively covers a sentence) and then identified which of the units served as arguments to each relation. Instead, as already noted, PDTB annotators were instructed to select the *minimal* clausal text span needed to interpret each of the two arguments to each explicit and implicit discourse connective.

We reported in ([Dinesh et al. 2005](#)) cases where the so-annotated arguments to discourse connectives diverged from syntactic constituents within the Penn TreeBank (PTB) — in particular, where instances of attribution (eg, “purchasing

[4] Independently of the PDTB, we also noticed that discourse relations could be conveyed by marked syntax – for example, the expression of a *CONDITIONAL* relation in the marked syntax of

(i) Had I known the Queen would be here, I would have dressed better.

or of a *CONTRAST* in the marked syntax of

(ii) The more food you eat, the more weight you'll gain.

Taking account of these examples will allow a more complete annotation of other corpora.

agents said” in Example (6) earlier and “analysts said” in Example (13)), headed non-restrictive relatives (eg, “which couldn’t be confirmed” in Example (13)), and headless non-restrictive relatives (eg, “led by Rep. Jack Brooks (D., Texas)” in Example (14)), all of which are constituents within the PTB’s syntactic analyses, are nevertheless not part of the discourse arguments of which they are syntactic constituents.

- (13) *ALTHOUGH* **traders rushed to buy futures contracts**, *many remained skeptical about the Brazilian development*, which couldn’t be confirmed, analysts said.
- (14) *Some Democrats*, led by Rep. Jack Brooks (D., Texas), *unsuccessfully opposed the measure because they fear that the fees may not fully make up for the budget cuts*.
BUT Justice Department and FTC officials said **they expect the filing fees to make up for the budget reductions and possibly exceed them**.

This suggests another way in which syntax and discourse may diverge, in addition to the lack of correspondence between coordination and subordination in syntax and in discourse noted by Blühdorn (2007): Material that is part of a constituent analysis in syntax may not be part of a constituent analysis in discourse. Also relevant here are examples such as (4) above, where constituency structure in syntax is standardly taken to be a tree, while in discourse, it appears to be a simple DAG (with the main clause a constituent of two distinct discourse connectives). Lee et al. (2006) presents an array of fairly complex constituency patterns of spans within and across sentences that serve as arguments to different connectives, as well as parts of sentences that don’t appear within the span of *any* connective, explicit or implicit. The result is that the PDTB provides only a partial but complexly-patterned cover of the corpus. Understanding what’s going on and what it implies for discourse structure (and possibly syntactic structure as well) is a challenge we’re currently trying to address.

[4] CONCLUSION

There is renewed interest in discourse structure, as more and larger annotated discourse corpora are becoming available for analysis. While there is still much to be gained from trying to extract as much as possible through machine learning methods based on superficial features of discourse that we believe we understand, there is also much to be gained from a deeper analysis of discourse structure that suggests new features that we are only now beginning to discover.

ACKNOWLEDGEMENTS

We would like to thank Nikhil Dinesh, Aravind Joshi, Mark Steedman and Bergljot Behrens for their comments on an earlier draft of this paper. The result of their suggestions is a much more coherent and, to our minds, more interesting paper.

REFERENCES

- Asher, N. and Lascarides, A. 2003. *Logics of Conversation*. Cambridge UK: Cambridge University Press.
- Blühndorn, H. 2007. Subordination and Coordination in Syntax, Semantics and Discourse: Evidence from the study of Connectives. In C. Fabricius-Hansen and W. Ramm (eds.), *'Subordination' versus 'Coordination' in Sentence and Text*, pages 59–89, John Benjamins.
- Bosma, W. 2004. Query-Based Summarization Using Rhetorical Structure Theory. In *Proceedings of 15th Annual Meeting of Computational Linguistics in the Netherlands (CLIN)*, pages 29–44.
- Carlson, L., Marcu, D. and Okurowski, M. E. 2003. Building a Discourse-Tagged Corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt & R. Smith (ed.), *Current Directions in Discourse and Dialogue*, New York: Kluwer.
- Creswell, C., Forbes, K., Miltsakaki, E., Prasad, R., Joshi, A. and Webber, B. 2004. The Predicate-Argument Structure of Discourse Connectives: A corpus-based study. In Tony McEnery Antonio Branco and Ruslan Mitkov (eds.), *Anaphora Processing: Linguistic, Cognitive and Computational Modeling*, John Benjamins.
- Di Eugenio, B., Moore, J. and Paolucci, M. 1997. Learning Features that Predict Cue Usage. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL97/EACL97)*, pages 80–87, Madrid, Spain.
- Dinesh, N., Lee, A., Miltsakaki, E., Prasad, R., Joshi, A. and Webber, B. 2005. Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives. In *ACL Workshop on Frontiers in Corpus Annotation*, Ann Arbor MI.
- Forbes, K. 2003. *Discourse Semantics of S-Modifying Adverbials*. Ph.D.thesis, Department of Linguistics, University of Pennsylvania.
- Forbes, K., Miltsakaki, E., Prasad, Rashmi, Sarkar, Anoop, Joshi, A. and Webber, B. 2003. D-LTAG System: Discourse Parsing with a Lexicalized Tree-Adjoining Grammar. *Journal of Logic, Language and Information* 12.
- Forbes-Riley, K., Webber, B. and Joshi, A. 2006. Computing Discourse Semantics: The Predicate-Argument Semantics of Discourse Connectives in D-LTAG. *Journal of Semantics* 23, 55–106.

- Grosz, B. and Sidner, C. 1986. Attention, Intention and the Structure of Discourse. *Computational Linguistics* 12(3), 175–204.
- Halliday, M. and Hasan, R. 1976. *Cohesion in English*. Longman.
- Kamp, H. and Reyle, U. 1993. *From Discourse to Logic*. Dordrecht NL: Kluwer.
- Kraemer, J. and Gibson, T. 2007. Ordering Constraints on Discourse Relations. In *Proceedings of the 20th Annual CUNY Conference on Human Sentence Processing*, page 116, La Jolla CA.
- Lascarides, Alex and Asher, Nicholas. 1993. Temporal Interpretation, Discourse Relations and Commonsense Entailment. *Linguistics and Philosophy* 16(5), 437–493.
- Lee, A., Prasad, R., Joshi, A., Dinesh, N. and Webber, B. 2006. Complexity of Dependencies in Discourse: Are dependencies in discourse more complex than in syntax? In *Proc. 5th Workshop on Treebanks and Linguistic Theory (TLT'06)*, Prague CZ.
- Mann, William and Thompson, Sandra. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3), 243–281.
- Marcu, D. 1996. Building up Rhetorical Structure Trees. In *Proceedings of AAAI-96*, pages 1069–1074, Portland OR.
- Marcu, D. 1998. Improving summarization through rhetorical parsing tuning. In *Proceedings of 6th Workshop on Very Large Corpora*, pages 206–215, Montreal, Canada.
- Marcu, D. 2000. *The theory and practice of discourse parsing and summarization*. MIT Press.
- Marcus, M., Santorini, B. and Marcinkiewicz, M. A. 1993. Building a Large Scale Annotated Corpus of English: The Penn TreeBank. *Computational Linguistics* 19, 313–330.
- Mellish, C., Knott, A., Oberlander, J. and O'Donnell, M. 1998. Experiments using stochastic search for text planning. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 98–107, Niagara-on-the-Lake, Canada.
- Miltsakaki, E., Prasad, R., Joshi, A. and Webber, B. 2004. Annotating Discourse Connectives and Their Arguments. In *NAACL/HLT Workshop on Frontiers in Corpus Annotation*, Boston.

- Miltsakaki, E., Robaldo, Livio, Lee, A. and Joshi, A. 2008. Sense Annotation in the Penn Discourse Treebank. In *Computational Linguistics and Intelligent Text Processing*, pages 275–286, Springer.
- Moore, J. 1995. *Participating in Explanatory Dialogues*. Cambridge MA: MIT Press.
- Moore, J. and Pollack, M. 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics* 18(4), 537–544.
- Moser, M. and Moore, J. 1996. Toward a Synthesis of Two Accounts of Discourse Structure. *Computational Linguistics* 22(3), 409–419.
- PDTB-Group. 2006. The Penn Discourse TreeBank 1.0 Annotation Manual. Technical Report IRCS 06-01, University of Pennsylvania, <http://www.seas.upenn.edu/~pdtb>.
- Polanyi, L. 1988. A Formal Model of the Structure of Discourse. *Journal of Pragmatics* 12, 601–638.
- Polanyi, L., Culy, C., van den Berg, M. H., Thione, G. L. and Ahn, D. 2004. Sentential Structure and Discourse Parsing. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 80–87, Barcelona, Spain.
- Polanyi, L. and van den Berg, M. H. 1996. Discourse Structure and Discourse Interpretation. In P. Dekker and M. Stokhof (eds.), *Proceedings of the Tenth Amsterdam Colloquium*, pages 113–131, University of Amsterdam.
- Prasad, R., Dinesh, N., Lee, A., Joshi, A. and Webber, B. 2007. Attribution and its Annotation in the Penn Discourse TreeBank. *TAL (Traitement Automatique des Langues)*.
- Prasad, R., Joshi, A., Dinesh, N., Lee, A., Miltsakaki, E. and Webber, B. 2005. The Penn Discourse TreeBank as a Resource for Natural Language Generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, Birmingham UK, slides available at <http://www.seas.upenn.edu/~pdtb>.
- Prasad, R., Miltsakaki, E., Joshi, A. and Webber, B. 2004. Annotation and Data Mining of the Penn Discourse TreeBank. In *ACL Workshop on Discourse Annotation*, pages 88–95, Barcelona, Spain.
- Ramm, W. and Fabricius-Hansen, C. 2005. Coordination and Discourse-structural Salience from a Cross-linguistic Perspective. In *Salience in Discourse: Multidisciplinary Approaches to Discourse*, pages 119–128, Munster, Germany: Stichting/Nodus.

- Schabes, Y. 1990. *Mathematical and Computational Aspects of Lexicalized Grammars*. Ph.D.thesis, Department of Computer and Information Science, University of Pennsylvania.
- Scott, D. and de Souza, C. S. 1990. Getting the Message Across in RST--based Text Generation. In Robert Dale, C. Mellish and Michael Zock (eds.), *Current Research in Natural Language Generation*, pages 47–73, London, England: Academic Press.
- Stede, M. 2004. The Potsdam Commentary Corpus. In *ACL Workshop on Discourse Annotation*, Barcelona, Spain.
- Stede, M. 2008. RST revisited: Disentangling nuclearity. In C. Fabricius-Hansen and W. Ramm (eds.), *'Subordination' versus 'Coordination' in Sentence and Text*, pages 33–59, Amsterdam: John Benjamins.
- Stede, M., Wiebe, J., Hajičová, E., Reese, B., Webber, B. and Wilson, T. 2007. Panel Session on Discourse Annotation. In *ACL Workshop on Language Annotation Workshop*, Prague.
- Steedman, M. 1996. *Surface Structure and Interpretation*. Cambridge MA: Linguistic Inquiry Monograph 30, MIT Press.
- Steedman, M. 2000. *The Syntactic Process*. Cambridge MA: MIT Press.
- Steedman, Mark. 2007. Surface Compositional Semantics of Intonation, submitted.
- Webber, B. 2004. D-LTAG: Extending Lexicalized TAG to Discourse. *Cognitive Science* 28, 751–779.
- Webber, B. 2005. A Short Introduction to the Penn Discourse TreeBank. In *Copenhagen Working Papers in Language and Speech Processing*.
- Webber, B. 2006. Accounting for Discourse Relations: Constituency and Dependency. In M. Dalrymple M. Butt and T. King (eds.), *Intelligent Linguistic Architectures*, pages 339–360, CSLI Publications.
- Webber, B., Stone, Matthew, Joshi, A. and Knott, Alistair. 2003. Anaphora and Discourse Structure. *Computational Linguistics* 29, 545–587.
- Wellner, B. and Pustejovsky, J. 2007. Automatically Identifying the Arguments to Discourse Connectives. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Prague CZ.
- Wiebe, J. 1993. Issues in Linguistic Segmentation. In *Workshop on Intentionality and Structure in Discourse Relations, Association for Computational Linguistics*, pages 148–151, Ohio StateUniversity.

Wolf, F. and Gibson, E. 2005. Representing Discourse Coherence: A Corpus-based Study. *Computational Linguistics* 31, 249–287.

XTAG-Group. 2001. A Lexicalized Tree Adjoining Grammar for English. Technical Report IRCS 01-03, University of Pennsylvania, see <ftp://ftp.cis.upenn.edu/pub/ircs/technical-reports/01-03>.

AUTHOR CONTACT INFORMATION

Bonnie Webber
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
United Kingdom
bonnie@inf.ed.ac.uk

Rashmi Prasad
University of Pennsylvania
Institute for Research in Cognitive Science
3401 Walnut Street, Suite 400A
Philadelphia PA 19104-6228
USA
rjprasad@seas.upenn.edu