Sentence-initial Discourse Connectives, Discourse Structure and Semantics

Bonnie Webber
University of Edinburgh
bonnie@inf.ed.ac.uk

Rashmi Prasad
University of Pennsylvania
rjprasad@seas.upenn.edu

This paper focusses on differences in the properties of *sentence-initial* (hereafter, s-initial) coordinating conjunctions (hereafter, CCs) and s-initial discourse adverbials, and what such differences imply for discourse structure and semantics. Data are mainly from the Penn Discourse TreeBank[1] (PDTB), which contains the annotation of discourse relations in 2304 articles of the Wall Street Journal corpus (Marcus *et al.*, 1993) in terms of *discourse connectives* and the minimal text spans that give rise to their arguments (Prasad *et al.*, 2008), as in Example 1:

(1) *Even so, according to Mr. Salmore, the ad was "devastating" because it raised questions about Mr. Courter's credibility.* <u>But</u> **it's building on a long tradition**. (0041)

The connective ("but") is underlined, the first of its two arguments, ARG1, is shown in italics and the second, ARG2, is shown in boldface. The number 0041 indicates that the example comes from subsection wsj_0041 of the corpus. All annotation has been done by two annotators and adjudicated by a third (often a committee) in cases of disagreement. Only where connectives serve to connect **clauses** (or possibly an event-denoting nominalization, in the case of ARG1), have they been annotated. The PDTB also annotates the sense of each connective and the attribution of both connectives and their arguments, but since neither plays a role in our presentation here, we omit further discussion.

As well as annotating explicit connectives, the PDTB annotates *implicit connectives*, which the annotators insert between paragraph-internal adjacent sentence not otherwise linked by an explicit connective, in order to express the inferred relation(s) between them. For example,

(2) *The projects already under construction will increase Las Vegas's supply of hotel rooms by 11,795, or nearly 20%, to 75,500.* **By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs**. (0994)

Here, the annotators take the connective "so" as implicitly connecting arguments found in the adjacent sentences, shown in italics (ARG1) and boldface (ARG2). Differences between this style of annotation and annotation in the RST corpus and in the GraphBank corpus are discussed in (Webber, 2006).

Here we focus on s-initial discourse connectives whose first argument comes from the preceding discourse (ie, CCs and discourse adverbials). Of interest is the location and form of ARG1, and what it is that cases where ARG1 is not immediately adjacent to the connective say about the intervening matieral. We present data on the CC "but" (as in Example 1) and on "so", which Huddleston and Pullum (2002) note has more similarities to CCs than it does differences. Both occur frequently in the *WSJ* corpus as s-initial clause connectors. While ARG1 of these connectives is always to their left and ARG2, to their right, ARG1 is not necessarily immediately adjacent to the connective.[2] Specifically, 20% of the 111 instances of s-initial "so" (22) have a non-adjacent ARG1, while 15.4% of s-initial "but" do so. While a similar pattern is found with discourse adverbials like "instead" and "nevertheless", does this mean that s-initial "but" and "so" behave not as CCs, but as adverbials?

Huddleston and Pullum (2002) describe *coordinators* (including CCs) as expressing "a relation between two or more elements of syntactically equal status". And in all instances of s-initial "so" and all but one instance of s-initial "but", we can show that ARG1 spans one or more clauses at a similar level of embedding as ARG2.[3] With s-initial discourse adverbials (which Huddleston and Pullum (2002) call "connective adverbials"), this is frequently not the case: Of the 61 s-initial instances of "instead" in the corpus, ARG1 spans text other than the main clause in 31 (51.7%), as in

(3) The tension was evident on Wednesday evening during Mr. Nixon's final banquet toast, normally an opportunity *for reciting platitudes about eternal friendship*. <u>Instead</u>, **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** (0093).

[1] available from the LDC, catalogue number LDC2008T05

[2] With *But*, figures are based on the first 410 of 2123 S-initial instances.

[3] Here we ignore attribution phrases such as "Mr. Dinkins argued", which are frequently not included in arguments to connectives (Dinesh *et al.*, 2005).

where ARG1 spans the gerund complement of an appositive NP. (The inclusion of "for" as part of ARG1 is simply a PDTB annotation convention.) Within a sentence, Huddleston and Pullum (2002) note that connective adverbials can link "non-coordinating elements". This clearly holds true in discourse as well.

Note, however, that when ARG1 of a s-initial "but" or "so" is non-adjacent to the connective, also of *syntactically equal status* is the intervening material. (This is not the case with discourse adverbials.) If both arguments of a s-initial CC are taken to be of "equal status" with respect to the discourse, then the intervening material might be serendipitous evidence for what Bluhdorn (2007) and others have called *subordination* in discourse, or what Mann and Thompson (1988) had in mind when they defined one argument of a rhetorical relation as a *satellite*, supporting the other argument (termed a *nucleus*). Subordination has not been directly annotated in the PDTB, but along with the exceptional instance of s-initial "but" mentioned above, it bears further analysis if we are to understand the annotation patterns in the PDTB, as well as discourse structure and semantics, more generally.

Also of interest is the patterning of s-initial discourse adverbials. If they can link non-coordinating elements in discourse (as well as within a sentence), then the coordinating elements (ie, the matrix sentence and a coordinating element in the previous discourse) can be linked by something else. Sometimes, this is an explicit connective, as with "so" in

(4) Long-winded people are boring, and writing full sentences is for chumps. So instead you have just SIX words to sum up the way you feel about Arsenal. (http://www.caughtoffside.com/?=pie, 6 May 2008)

More often, an *implicit connective* can be seen to link the matrix sentence of the connective and ARG2 to the previous discourse, as in

(5) *No price for the new shares has been set.* Instead, **the companies will leave it up to the marketplace to decide**. (0018)

where an implicit "and" is justified. That this implicit connective is not always "and" can be seen in such minimal pairs as

(6) a. *There wasn't any bread.* Instead **we ate crackers**.

b. *We wanted to eat bread.* Instead **we ate crackers**.

where "so" is an appropriate implicit connective for (a), while "but" is an appropriate implicit connective for (b). Although implicit connectives have not been annotated in the PDTB in cases of s-initial discourse adverbials, it is worth doing so, in order to more fully capture discourse semantics and structure.

References

Bluhdorn, H. (2007). Subordination and coordination in syntax, semantics and discourse In C. Fabricius-Hansen and W. Ramm, editors, *Subordination versus Coordination in Sentence and Text — from a cross-linguistic perspective*. John Benjamins.

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.

Huddleston, R. and Pullum, G. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge UK.

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

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.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings, 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Webber, B. (2006). Accounting for discourse relations: Constituency and dependency. In M. D. M. Butt and T. King, editors, *Intelligent Linguistic Architectures*, pages 339–360. CSLI Publications.