Anaphoric arguments of discourse connectives: Semantic properties of antecedents versus non-antecedents

Eleni Miltsakaki, Cassandre Creswell, Katherine Forbes, Aravind Joshi,
Institute of Research in Cognitive Science
University of Pennsylvania
elenimi@linc.cis.upenn.edu
creswell@babel.ling.upenn.edu
forbesk@linc.cis.upenn.edu
joshi@linc.cis.upenn.edu

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

Abstract

We have argued extensively in prior work that discourse connectives can be analyzed as encoding predicate-argument relations whose arguments derived from the interpretation of discourse units. All adverbial connectives we have analyzed to date have expressed binary relations. But they are special in taking one of their two arguments structurally, and the other, anaphorically. As such, interpreting adverbial discourse connectives can be understood as a problem of anaphora resolution. In this paper we study the S-modifying adverbial connective "instead" and what, in the context, does and does not serve as antecedent for its anaphoric argument. This work extends earlier work investigating syntactic patterns of anaphoric arguments across a range of adverbial discourse connectives and the reliability with which these arguments can be annotated. The current work establishes, for 100 successive corpus instances of "instead", lexico-syntactic features of the antecedents of their anaphoric arguments that can be automatically annotated and therefore used to distinguish actual antecedents from potential competitors in the context.

1 Introduction

Discourse relations can be lexicalized in at least two ways – with subordinate/coordinate conjunctions and with adverbial phrases,[1] as in:

(1) **Subordinate conjunction.** *Although* Mr. Hastings had been acquitted by a jury, lawmakers handling the prosecution in Congress had argued that the purpose of impeachment isn't to punish an individual.

(2) **Coordinate conjunction.** The Berkeley police don't have any leads *but* doubt the crime was driven by a passion for sweets.

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

Both types of connectives can be analyzed as encoding predicate-argument relations whose arguments derive from the interpretation of discourse units (Webber and Joshi, 1998). With subordinate or coordinate conjunction, those discourse units are the ones structurally joined by the conjunction, thus enabling the semantics of the relation to be built compositionally via well-understood mappings of syntax to semantics (Webber et al., 1999). We call subordinate and coordinate conjunctions *structural connectives*. For example, the structural connective *although* in (4) expresses a concessive relation between the two eventualities, P = RARELY EAT (SALLY, MEAT) and Q = ENJOYS (SALLY, CHEESEBURGER).

(4) Although Sally rarely eats meat,
she enjoys an occasional bacon cheeseburger.

All the adverbial connectives we have analyzed to date express binary predicate-argument relations. They differ from structural connectives in only getting one of their two arguments structurally – the one they get from their matrix clause. With respect to their other argument, we have argued extensively (Webber and Joshi, 1998; Webber et al., 1999; Webber et al., 2003) that adverbial connectives behave like common discourse anaphors (pronouns and NPs), obtaining this argument from the discourse context. The problem of interpreting adverbial connectives with respect to the discourse can thus be reformulated as an *anaphor resolution problem*, and, for this rea-

[1] Discourse relations can also be lexicalized with a *null* connective as in: 'You should not lend Tom any books. He never returns them'. While we have included null connectives in previous studies, they are not discussed in this paper.

son, we often call adverbial connectives *anaphoric connectives.* For example, if (4) were followed by

(5) Otherwise, she would pine away for lack of grease.

the adverbial connective *otherwise* conveys a conditional relation between the complement of Q = ENJOYS (SALLY, CHEESEBURGER) and R = PINE AWAY (SALLY).

From both a theoretical perspective (Gundel et al., 1993; Prince, 1981; Walker and Prince, 1996; Prince, 1999) and an empirical perspective, it is clear that different discourse anaphors (e.g., third person pronouns, definite NPs, demonstrative pronouns, demonstrative NPs, "other" NPS, etc.) display different properties with respect to where and what in the discourse context they can draw their referents from. For particular anaphors or sets of anaphors, empirical studies can help elucidate what those properties are.

In this paper, we report on an empirical study of the surprisingly interesting adverbial connective "instead". "Instead" occurs in two forms: (1) as a bare adverbial, and (2) with an "of" PP modifier. In the latter form, it can be found with every type of phrase, including NPs (Example 6), AdjPs (Example 7), and PPs (Example 8):

(6) John ate an apple instead of a pear.

(7) John chose a bright yellow instead of a dull blue shirt.

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

In this form, both arguments of "instead" can be derived structurally: the first, from the phrase it modifiers (e.g., "an apple" in Example 6) and the second, from the object of its "of" PP (e.g., "a pear" in the same example). Semantically, the "instead of" phrase conveys that its second argument (here, the "of" PP) is a salient but unchosen *alternative* to its first argument, with respect to the given predication. The notion of salient but unchosen or unrealized *alternatives* is basic to the interpretation of "instead" in both its modified and bare forms.

As a bare adverbial, "instead" gets its second argument *anaphorically*, from the discourse context. As before, this argument corresponds to a salient but unchosen or unrealized alternative. But not every discourse context provides salient alternatives for the anaphoric argument of "instead" to resolve with, and therefore its use is not always licensed – e.g.

(9) John ate an apple. #Instead he wanted a pear.

(10) John wanted to eat a pear. Instead he ate an apple.

(11) John won't eat fruit. Instead, he eats only candy bars and potato chips.

To better understand "instead" as a discourse connective, we carried out an empirical study of its discourse context and the properties of what, in that context, did and did not serve as an antecedent for its anaphoric argument. The results will help in the development of an anaphor resolution mechanism for "instead" and a methodology for developing anaphor resolution mechanisms for other anaphoric connectives.

2 Previous Work

Our first empirical work in this area (Creswell et al., 2002) was aimed at verifying the distinction between structural and anaphoric connectives that we had argued for on theoretical grounds. We described a preliminary corpus annotation effort for nine discourse connectives. The results indicate that classes of connectives display distinctive resolution patterns, as do individual connectives. The preliminary annotation included mainly *surface syntactic features* such as the location and size of the argument, its clausal characteristics and the location of the connective. Consistent with expected attentional constraints, most of the studied connectives had a strong tendency for their left argument to be identified locally (in the structural sense) – either in the immediately preceding sentence or in immediately preceding sequence of sentences, in most cases the preceding paragraph. Most notably, it was observed that *so* always takes a sentence or a sequence of sentences as its left argument, indicating that it might be treated as a structural connective. *In addition*, *yet*, *moreover*, *as a result* and *also*, tend to take their left argument locally but they demonstrate a larger syntactic variety of potential arguments such as subordinate clauses or phrasal constituents. Finally, *so*, *nevertheless* and *moreover* are more likely to take larger discourse segments as arguments.

3 Corpus Annotation Study: "Instead"

Our annotation study of the anaphoric connective *instead* has two parts: (1) annotation of the antecedent of its anaphoric argument and of lexico-syntactic features of that antecedent that could correlate with semantic properties suggestive of salient *alternatives* that could be (but haven't been) realized or chosen; and (2) annotation of clauses in close proximity to this antecedent which could potentially serve as distractors or *competing antecedents* (cf. Section 3.2).

The purpose of the first part of the study was to establish whether every true antecedent of *instead* could be characterized in terms of lexico-syntactic features that could be automatically annotated. The purpose of the second part was to establish whether *competing alternatives* displayed such features. If they didn't, then the absense of any such features on previous clauses close to bare *instead* could be used to *reject true negatives*, and the presense of such features or feature sets could be used to *strongly suggest a true positive*.

3.1 Annotation of competing antecedents

We examined 100 successive instances of sentence initial *instead*, in each case (a) identifying the text containing the antecedent of its anaphoric argument; (b) computing inter-annotator agreement; and (c) annotating lexico-syntactic features of the antecedent. We then quantified the frequency of appearance of these features in the identified arguments.

The features we chose to annotate were ones present in instances of *instead* that we had previously collected serendipitously: clausal negation, presence of a monotone-decreasing quantifiers (e.g., *few*, *seldom*), presence of a modal auxiliary, presence of conditionality, and verb type. In addition, some of our serendipitously collected examples showed the antecedent of *instead* embedded in a higher clause, and thereby *not* part of the assertions of the sentence, as in clause (12), which neither entails, presupposes, nor implicates clause (13). So we also annotated whether or not an antecedent was embedded in some higher clause.

(12) John wanted to eat a pear.

(13) John ate a pear.

Table 1 contains the complete set of features used in annotating the antecedent of the anaphoric argument of *instead*.

Feature	Abbreviation
Verbal negation	(Verbal neg.)
Subject negation	(Subj. neg.)
Object negation	(Obj. neg.)
Monotone decreasing quantifier	(MDQ)
Modal auxiliary	(Modal)
Conditional sentence	(Condit.)
Embedded antecedent	(Embed.)

Table 1: Set of annotation features

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

Table 2: Results from antecedent annotation of the anaphoric argument

3.1.1 Results from the antecedent annotation

Table 2 shows the results of this annotation for 97 out of the 100 tokens in the original set. In the remaining three cases the annotators did not agree on the argument of *instead*, and these cases were excluded from further analysis.

Antecedents could display zero or more of the features from the set given in Table 2 – for example, both a negative subject and a modal auxiliary, or no value for "object negation" if the verb in the antecedent clause is transitive. Three things stand out: (1) the presence of negation on the verb or one of its arguments, (2) the presence of a modal auxiliary, and (3) the presence of a higher verb. (We discuss this last feature and its significance in Section 3.2).

In the majority of tokens (65 of 97 cases), at least one of the first six features in Table 2 (i.e. all

features but EMBED) was present in the antecedent of the anaphoric argument of *instead*. In an additional 27% of tokens, the semantics of either the verbal predicate in the antecedent itself or the verbal predicate embedding the antecedent admits alternative situations or events (e.g., expect, want, deny etc.), such as *demand* in (14), which embeds the antecdent clause *that he surrender*. We will see in the next section, that the frequency of these features in antecedent clauses is significantly greater than their frequency in clauses which do not serve as antecedents of *instead*.[2]

(14) Arriving at daybreak , they found Julio in his corral and **demanded** *that he surrender. Instead, he whirled and ran to his house for a gun,* forcing them to kill him , Cook reported .

In sum, for a total of 94% of tokens, we were able to characterize features of the arguments that could be automatically extracted from existing annotations and used to help resolve these anaphoric arguments. In the remaining cases, the annotated features were absent, meaning that the set of conditioning features is incomplete.

3.2 Annotation of competing antecedents

For the annotation of competing antecedents, we defined *competing antecedent* as follows: any finite or non-finite clause contained in the sentence which contains the antecedent of *instead* or that intervenes between the antecedent and the sentence containing *instead*.

We adopt the traditional definition of sentence, which contains a single main verb and all its associated finite or non-finite clauses including relative and adverbial clauses. We have also classified as 'sentence' instances with two main verbs in cases of VP coordination, i.e., when the subject of the second verb is omitted. While other definitions of "competing antecedents" are plausible, our definition takes advantage of earlier results which show that in most cases the antecedent of the anaphoric argument of *instead* is contained within the immediately preceding sentence or shortly before it.

[2]In the case of the feature EMBED, it is the semantics of the embedding verb, not just its syntactic properties that make it a conditioning feature for the presence of an antecedent argument. However, as will be seen below, the frequency of embedding in actual vs. potential antecedents indicates that even disregarding the identity of the embedding verb, this is a useful property for identifying actual antecedents.

The same set of features used in annotating the anaphoric argument was also used in annotating *competing antecedents*. We made this choice as a preliminary step in building an anaphora resolution algorithm. Our primary goal in annotating competing antecedents with the same set of features was to evaluate their strength in distinguishing arguments from non-arguments in a well-defined syntactic locality.

3.2.1 Results

For the set of 97 tokens of *instead* extracted in the first part of this study, we identified 169 tokens of 'competing antecedents'. Table 3 shows the results of the annotation. Overall, comparing Tables 2 and 3, two things stand out:

1. Negation of the verb or one of its arguments is much more common in the antecedent of *instead* than in potentially competing antecedents – 52/97 times ($\approx$ 53%) versus 35/169 times ($\approx$ 20%).

2. The antecedent of the anaphoric argument of *instead* is much more frequently embedded in a higher verb than is a potentially competing antecedent – 57/97 times ($\approx$ 59%) vs 14/169 times ($\approx$ 8%).

In our annotated example set, we do not have enough instances of monotonically decreasing quantifiers, modal auxiliaries or conditionals to say whether their co-occurrence with the antecedent of the anaphoric argument of *instead* is significantly different from their co-occurrence with potentially competing antecedents.

Features	YES (of 169)	No (of 169)
Verbal neg.	21 (12%)	148 (88%)
Subj. neg.	8 (5%)	161 (95%)
Obj. neg.	6 (4%)	139 (82%)
MDQ	0 (0%)	169 (100%)
Modal	17 (10%)	152 (90%)
Condit.	0 (0%)	169 (100%)
Emb.	14 (8%)	155 (91%)

Table 3: Results from feature annotation of competing antecedents including higher verbs

The difference between these sets is significant. In the first case, the embedded clause may be *desired* ("want", "advise", "insist") or *(un)expected* ("expect", "doubt"), *described* ("tell", "say"), etc. but is not *asserted* or *presupposed* to hold now or have held before or to hold in the future. This makes other alternatives that could hold both possible and salient, one of which is the structural argument of *instead*. Given the current set of seven features here, a very simplistic resolution algorithm based on implementing these features directly (i.e. if *any* of the token's features' values is Y, then the token should be marked as ANTECEDENT=Y) would have very good recall, but poor precision. Among other possible improvements, consideration of the features of the structural argument of *instead*, along with the features of the potential antecedent candidate, could presumably decrease the incidence of these false positives.

What is not obvious from Table 2 is the *nature* of the higher verbs that actual antecedents and potentially competing antecedents occur with.

Features	YES (out of 39)	NO
Verbal neg.	4 (10%)	35 (90%)
Subj. neg.	1 (2%)	38 (98%)
Obj. neg.	2 (5%)	20 (51%)
MDQ	0 (0%)	39 (100%)
Modal	2 (5%)	37 (95%)
Condit.	1 (1%)	38 (99%)
Emb	2 (5%)	37 (95%)

Table 4: Results from feature annotation of higher verbs

Table 4 shows that the set of features that appears to be successful for distinguishing between actual and competing antecedents is not equally useful for distinguishing between verbs that embed the antecedent and competing antecedents. A better characterization of the class of higher verbs will be achieved by looking at differences of the semantic properties of higher verbs that embed antecedents from those that do not. In the case of antecedents, higher verbs included *insist, abandon, doubt, expect, tell, say, concede, want, be appropriate*), while higher verbs of potentially competing antecedents included factive verbs such as *know*. A clause with a factive verb can give rise to salient alternatives, but not to alternatives to the embedded clause because factive verbs presuppose its truth, as in (15). The continutation with *instead* is possible in (16) but this is because of the presence of negation in the higher clause.

(15) John regretted eating 12 bananas. *Instead ...

(16) John didn't regret that eating 12 bananas. Instead he was happy. (*Instead* possible because of the negation)

Note that the *antecedent* of the anaphoric argument of *instead* is not the same as the abstract object that serves as that argument (Webber et al., 2003). Deriving arguments from antecedents may require inference. However, as with resolving discourse deixis (Webber, 1991; Eckert and Strube, 2001; Byron, 2002), properties of the matrix clause containing the anaphor (here, *instead*) can constrain the inference process. Thus in Example 17, the fact that the anaphoric argument of *instead* is an alternative to what Valhi and affiliates will do with their Lockheed holdings, allows one to infer from the (bolded) antecedent, that that argument is (roughly) Valhi and affiliates doing with respect to Lockheed what the article said it would.

(17) Valhi Inc., another of Mr. Simmons' companies, responded to an article Monday in The Wall Street Journal, which credited a story in the Sunday Los Angeles Daily News. Valhi said **the articles didn't accurately reflect Valhi and its affiliates' intentions toward Lockheed**. Instead, Valhi said, they may increase, decrease or retain their Lockheed holdings , depending on a number of conditions.

4 Discussion

The set of annotated antecedents of the anaphoric arguments of *instead* contained cases in which no feature from our set was present. These cases are particularly interesting as they highlight the complex nature of the lexico-syntactic realization of semantics that give rise to alternatives. In (18), for example, annotators agreed that the antecedent of *instead* was the phrase shown in boldface. However, this phrase has none of our annotated features and the predicate 'recite' is not one that appears to give rise to alternatives.

(18) 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 kil led hundreds, and perhaps thousands, of demonstrators.

What appears to trigger alternatives are the phrases "normally" and "an opportunity", either individually or together. The fact that (19) and (20) are comparable to (21) and (22), suggests that the range of lexical items triggering alternatives is larger than negation and monotone decreasing quantifiers, modality and certain classes of verbal predicates, and moreover, does not correspond to any previously defined set of linguistic elements.

(19) I had the opportunity to buy a cheap used car. Instead, I bought a scooter.

(20) This event was an opportunity for John to make amends. Instead, he caused more trouble.

(21) I wanted to buy a car. Instead I bought a scooter.

(22) John could have made amends. Instead he caused more trouble.

5 Conclusion

In earlier work we argued that adverbial connectives take one argument structurally and one anaphorically. In this paper, we looked at the lexico-syntactic realization of the antecedent of the anaphoric argument of *instead*. For anaphora resolution, the advantage of identifying lexico-syntactic realizations of the relevant semantic features is that such features can easily be extracted automatically from available sources such as the syntactic annotation of the Penn Treebank corpus and the semantic annotation of the Penn PropBank corpus. In future work, we plan to conduct a large scale corpus annotation project on top of the Penn Treebank and Penn PropBank in order to study (a) the semantic properties of higher verbs embedding the antecedent, (b) the relationship between the structural and anaphoric argument of *instead*, and (c) additional semantic properties of the arguments of *instead* that will be useful in identifying the antecedent of the anaphoric argument.

References

Donna Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting, Association for Computational Linguistics*, pages 80–87, University of Pennsylvania.

Cassandre Creswell, Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2002. The discourse anaphoric properties of connectives. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002), Lisbon, Portugal*, pages 45–50. Edições Colibri. (First four authors in alphabetical order).

Miriam Eckert and Micahel Strube. 2001. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*.

Jeanette Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.

Ellen Prince, 1981. *Radical Pragmatics*, chapter Toward a Taxonomy of Given-New Information, pages 223–255. NY: Academic Press.

Ellen Prince, 1999. *Focus: Linguistic, Cognitive, and Computational Perspectives*, chapter Subject Pro-Drop in Yiddish, pages 82–101. Cambrige University Press.

Marilyn Walker and Ellen Prince. 1996. A bilateral approach to givenness: A hearer-status algorithm and a Centering algorithm. In T. Fretheim and J. Gundel, editors, *Reference and Referent Accessibility*, pages 291–306. Amsterdam: John Benjamins.

Bonnie Webber and Aravind Joshi. 1998. Anchoring a lexicalized tree adjoining grammar for discourse. In *ACL/COLING Workshop on Discourse Relations and Discourse Markers, Montreal*, pages 8–92. Montreal, Canada.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse relations: A structural and presuppositional account using lexicalized TAG. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Maryland*, pages 41–48. College Park MD.

Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*.

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