Filling in the Blanks in Understanding Discourse Adverbials: Consistency, Conflict, and Context-Dependence in a Crowdsourced Elicitation Task

Hannah Rohde[*] **Anna Dickinson**[*] **Nathan Schneider**[*,†]
Christopher N. L. Clark[*] **Annie Louis**[‡] **Bonnie Webber**[*]

[*]University of Edinburgh
Edinburgh, UK

[†]Georgetown University
Washington, DC, USA

[‡]University of Essex
Colchester, UK

{hannah.rohde, bonnie.webber}@ed.ac.uk,
{anna.y.dickinson, chrisclark272}@gmail.com,
nschneid@inf.ed.ac.uk, aplouis@essex.ac.uk

Abstract

The semantic relationship between a sentence and its context may be marked explicitly, or left to inference. Rohde et al. (2015) showed that, contrary to common assumptions, this isn't *exclusive or*: a conjunction can often be inferred alongside an explicit discourse adverbial. Here we broaden the investigation to a larger set of 20 discourse adverbials by eliciting ≈28K conjunction completions via crowdsourcing. Our data replicate and extend Rohde et al.'s findings that discourse adverbials do indeed license inferred conjunctions. Further, the diverse patterns observed for the adverbials include cases in which more than one valid connection can be inferred, each one endorsed by a substantial number of participants; such differences in annotation might otherwise be written off as annotator error or bias, or just a low level of inter-annotator agreement. These results will inform future discourse annotation endeavors by revealing where it is necessary to entertain implicit relations and elicit several judgments to fully characterize discourse relationships.

1 Introduction

Existing work highlights the importance of understanding discourse relations in context, showing a range of phenomena that are sensitive to the semantic connection that holds between two spans of discourse (Hirschberg and Litman, 1987; Kehler and Rohde, 2013). Such connections can be made explicit in text via an overt connective or marked syntax; otherwise they must be inferred. Various contextual cues have been identified that guide the establishment of discourse relations (Hirschberg and Litman, 1987; Kehler, 2002; Webber, 2013).

When it comes to producing resources annotated with discourse relations—e.g., the Penn Discourse Treebank (PDTB; Prasad et al., 2008)—it is commonly assumed that at most a *single* discourse relation holds between two spans of discourse. It may not be simple to identify or infer that relation, but once achieved, the task is taken to be done. But properties of the discourse adverbial **instead** (Webber, 2013) have challenged this assumption. In particular, sentence-initial **instead** supports the inference of another discourse relation, with the specific relation depending on properties of the spans. This can be seen through what coordinating conjunction makes the relation explicit—compare:

(1) I planned to make lasagna. Instead I made hamburgers.
⇒ But instead I made hamburgers

(2) I don't know how to make lasagna. Instead I made hamburgers.
⇒ So instead I made hamburgers

(3) Surprisingly, they ignored the lasagna. Instead they just ate the salad.
⇒ And instead they just ate the salad

While this means that full annotation of **instead** requires asking annotators what additional relation they infer (besides that associated with **instead** itself), one still needs to ask:

- For clauses starting with discourse adverbials other than **instead**, is the relation signalled by the adverbial all there is, or can an additional relation be inferred with the previous text? In the former case, no additional annotation is required; in the latter, it is.

- If another relation can be inferred, can it be inferred deterministically based on the adverbial alone? If so, no additional work is required, as the relation can be annotated automatically.
- If it can't be inferred based on the discourse adverbial alone (as in the case of **instead**), how should an annotator figure out what it is?
- Could there be different ways of framing the inferred relation, such that annotators may disagree as to its identity, but all still be correct?

This paper addresses these questions using crowdsourced data elicited on 969 passages involving twenty discourse adverbials. We describe our methodology, what we have so far been able to learn, and how *inter-annotator disagreements* have led us to look more deeply into the judgments and what conclusions we can draw from them. Our results demonstrate that inter-annotator disagreement is informative, and need not be treated as annotator bias, inattention, or noise.

2 Background

The current work should be seen against the background of two research areas: Research on multiple co-occurring connectives and research on acquiring useful linguistic judgments from a large number of annotators, whether by crowdsourcing or in-house.

In the PDTB, all explicit connectives in a sentence were separately annotated. Then, *if and only if* a sentence lacked an explicit inter-sentential connective linking it to the previous context, annotators were asked to infer and annotate its relation, if any, to the previous sentence. This reflected the common assumption, noted earlier, that the situation is "either/or" – if a discourse relation is marked, there is nothing to infer.

With respect to research on explicit multiple co-occurring connectives, over 15 years ago, Webber et al. (1999) used them to argue that discourse spans could be related by both adjacency relations and anaphoric relations. Similary, in the context of Catalan and Spanish oral narrative, Cuenca and Marín (2009) used them to argue for different patterns and degrees of discourse cohesion. Oates (2000) considered how multiple discourse connectives should be used in Natural Language Generation, noting that the order in which they occur correlates with the hierarchy of discourse connectives presented in (Knott, 1996), while Fraser (2013) offers an account of the order in which multiple *contrastive* connectives co-occur, in terms of what he calls *general contrastive* discourse markers and *specific contrastive* discourse markers. For Turkish, Zeyrek (2014) has described patterns of multiple co-occuring connectives that signal *contrastive* and/or *concessive* relations.

These efforts have all been directed at explaining the existence of multiple explicit connectives and how they pattern. Closer to the focus of the current paper is work by Rohde et al. (2015), in which judgments were crowdsourced on four adverbials: **after all**, **in fact**, **in general** and **instead**. Rohde et al. found that, given one of these discourse adverbials, naïve participants identified an operative discourse relation—via a conjunction whose presence they endorsed alongside the discourse adverbial. They did so reliably both for explicit passages in which the author's explicit pre-adverbial conjunction had been elided and for implicit passages in which the adverbial originally appeared alone. For Rohde et al.'s four adverbials, the inferred relation could not be predicted entirely on the basis of the adverbial alone. The current study extends Rohde et al.'s work to a larger set of adverbials. We focus on participant judgments on implicit passages since such cases are left largely untreated by existing annotation endeavors as well as current formal accounts.

The other research area that forms the background to the current work is research on acquiring linguistic judgments from a large number of annotators, whether by crowdsourcing or in-house. Here, research has addressed either identifying and correcting for problems arising from judgments from large numbers of unknown, possibly biased and/or inattentive annotators (Hovy et al., 2013; Passonneau and Carpenter, 2014), or identifying benefits that arise from having a large number of annotators (Artstein and Poesio, 2005, 2008). Work in the former area attempts to eliminate judgments that should be treated as noise, while the latter work shows that annotator bias decreases with the number of annotators.

In related research, Poesio and Artstein (2005) reflect on the "true ambiguity" of some pronoun tokens and how the presence of these distinct co-present viable interpretations can be brought to light via a sufficiently large number of annotators. In one example they cite, a boxcar has been attached to a train engine. The next sentence specified what should then be done. Over half their participants interpreted the pronoun *it* in this next

sentence as referring to the boxcar, while others interpreted it to refer to the engine. But the situation associated with these two different interpretations was the same in both cases, since the engine and boxcar had effectively become a single moving, functioning unit. This ambiguity would not necessarily have been made apparent or taken to be as significant without the large number of participants.

Lastly, there is new work (Scholman et al., 2016) that tests naïve annotators' ability to infer discourse relations, specifically to distinguish four dimensions along which relations are posited to vary. Their work targets annotator agreement, and shows consistency comparable with expert annotators for two of the four posited dimensions. Unlike their task, which asked participants to make a decision about abstract semantic features, our methodology involves asking participants to consider whether a short passage with an explicit conjunction is a paraphrase of one without that conjunction. Crucially, we will avoid the assumption that there is a single correct answer.

3 Crowdsourcing judgments on discourse adverbials: Methodology

Here we extend the crowdsourcing approach of Rohde et al. (2015) to a larger dataset with many more adverbials. The goal is to learn from participants' endorsements of particular conjunction–adverbial combinations in naturally occurring passages, as to whether additional annotation will be needed.

3.1 Participants

We recruited 28 participants from Amazon Mechanical Turk. All were native English speakers and were paid $88 each for their participation. These 28 individuals were selected from a larger pool who participated in a pre-trial involving 50 annotations. The pre-trial allowed us to identify participants who understood the task, whose responses were in line with the group average, who did not overuse NONE, and who were not outliers in speed.

3.2 Materials

The target passages that participants read were selected from the NY Times Annotated Corpus (Sandhaus, 2008). Passages varied from 9 to 122 words (minimally a sentence and maximally, a short paragraph). They were chosen to be comprehensible as stand-alone excerpts. Each target passage consisted (minimally) of two spans of text, the second beginning with a discourse adverbial, as in examples (1)–(3) and the sample materials shown in (4)–(5).

(4) "Nervous? No, my leg's not shaking," said Griffey, who caused everyone to laugh / _____ indeed his right foot was shaking.

(5) Sellers are usually happy, too / _____ after all / they are the ones leaving with money.

In example (4)'s original form, the author had included an explicit conjunction (*because*). In example (5), the original text contained only the adverbial, meaning that a discourse relation conveyed by a conjunction would have been implicit. Punctuation adjacent to the adverbial was replaced with a slash.

Each passage contained one of the following discourse adverbials after the gap: **actually**, **after all**, **first of all**, **for example**, **for instance**, **however**, **in fact**, **in general**, **in other words**, **indeed**, **instead**, **nevertheless**, **nonetheless**, **on the one hand**, **on the other hand**, **otherwise**, **specifically**, **then**, **therefore**, and **thus**. These represent a sampling of high-frequency adverbials, which belong to a variety of semantic classes and which showed a range of conjunction co-occurrence patterns in counts extracted from the Google Books Ngram Corpus (Michel et al., 2011; Lin et al., 2012).

Half the target passages originally contained a conjunction before the adverbial. For those *explicit passages*, we excised the conjunction and replaced it with a gap. For excerpts that were originally *implicit passages*, we simply inserted a gap before the adverbial. For each of the 20 adverbials, participants saw 25 explicit passages and 25 implicit passages, with the exception of **however**, which appeared in 25 implicit passages and 1 explicit passage (due to the rarity of conjunctions that naturally occur directly before **however**).

The distribution of original (author-chosen) conjunctions in the explicit passages reflected the distribution observed in Google n-gram counts of each adverbials with each of the conjunctions AND, BECAUSE, BUT, OR, SO. These 5 conjunctions appeared in a list of possible response options for participants.

With 20 adverbials and 50 passages per adverbial (26 passages for **however**), this yields a set of 976 passages. Due to presentation errors in 7 passages, the dataset for analysis consists of participant responses to 969 unique passages. The experiment

also included 32 catch trials, which were used to check that participants were paying attention and using the experimental interface correctly. The catch trials contained well-known quotes and expressions that had a 'correct' conjunction (e.g., *you can lead a horse to water _____ you can't make it drink*). Some of the catch trials expected the response BEFORE, so this was included as a 6th option in the list of possible conjunction responses.

How much can we learn from participants' selection of a conjunction? All six conjunctions we use are relatively unambiguous: In the PDTB (Prasad et al., 2008), each has a different main sense that it is associated with >90% of the time.[1] More to the point, while 5.9% of the explicit tokens of **and** were assigned a *result* sense, of the 1272 tokens where AND was inserted as an implicit connective, none were labelled with the inferred sense *result*. (97% of the time, when AND was inserted as an implicit connective, it was with an inferred sense of *conjunction* or *list*, as with explicit tokens of AND.) As such, there are grounds for believing that the experiment targeted the participants' inferred relation through choosing a conjunction that realizes it, even if the sense is only a coarse one.

3.3 Procedure

All participants saw all passages. Participants were instructed to fill in the gap with the word of their choice (from the six conjunctions AND, BECAUSE, BEFORE, BUT, OR, SO) that "best reflects the **meaning** of the connection" between the spans. They also had the option of choosing either NONE AT ALL (if they felt that no conjunction was possible) or OTHER WORD OR PHRASE (if they felt that only some option other than the six presented conjunctions was appropriate). The instructions were followed by three practice items.

During pilot testing, it emerged that participants sometimes chose NONE AT ALL when it sounded more fluent and less awkward to them than did an explicit conjunction. To avoid this, we explicitly instructed participants to choose the conjunction that best conveyed the sense of the connection, "even if the resulting text sounds awkward", but then offered them the opportunity to record whether or not they would in fact use the chosen conjunction in that context (recording "I could say it this way" or "It sounds strange here").

[1] OR has the sense *Disjunction* 86.7%, since it is labelled as *Conjunction* when it is in a negative context.

To avoid order effects, passages were pseudo-randomised: Participants never encountered more than three of the same adverbial in a row, and for explicit passages, they never saw excerpts whose original (author-chosen) conjunction was the same more than three times in a row. Also randomized was the list of possible conjunctions from which participants selected their response: The list appeared in a different order for each participant. Cursory examination of a sample of the data fails to show any obvious bias from the order in which the choices were presented.

The task was completed over several weeks. Participants worked at a rate of roughly 85 tokens per hour (making the hourly rate roughly $8/hour). They were not permitted to do more than 100 tokens per day.

4 Results

4.1 Issues addressed by the results

Our crowdsourced data can be used to answer distinct questions: Responses on explicit passages (§4.2) can be used to test whether untrained participants *can* do the task and deliver useful information. Given that the answer is found to be 'yes', responses on implicit passages (§4.3) can be used to answer our fundamental research questions: (1) Do inferrable discourse relations hold in implicit passages containing only a discourse adverbial, and (2) how can individual adverbials best be characterized with respect to inferrable discourse relations?

Assuming that the first question is answered in the affirmative (as was shown by Webber (2013) for **instead**), the two strongest answers to the latter question would be either:

i. **Uniformity across adverbials:** All adverbials co-occur with the same preferred conjunction.

ii. **Uniformity per adverbial:** Each adverbial has a single preferred conjunction, not necessarily the same across adverbials but possibly predictable from the semantic class of the adverbial.

If either was the case, it would be straightforward to obtain additional annotation of discourse adverbials. However, §4.3 will show that conjunctions preferred by participants are neither uniform across adverbials (contra (i)) nor uniform across passages for a particular adverbial (contra (ii)). We can nevertheless use these two types of variability to characterize the adverbials in this study.

§5 will then show that systematic variability in the responses of our untrained annotators reveals cases in which multiple interpretations are inferable—an outcome that, as in (Poesio and Artstein, 2005), only presents itself with the use of multiple annotators. We discuss implications for large-scale annotation frameworks and methods.

4.2 Responses for explicit passages

For each of the 20 adverbials in our study, we elicited responses for 25 *explicit* passages, where the original sentence contained an adverbial preceded by a conjunction. (As already noted, for **however** we elicited responses for only one explicit passage; also, in the case of four explicit passages containing other adverbials, there were errors in presentation.) With 28 participants who saw all of these explicit passages, we have 13,216 analyzable data points.

The results show that conjunctions selected by authors in original texts are indeed recoverable: More than half the time (57%), participants selected the conjunction that the author had used. Moreover, it has been noted that the conjunction AND provides a less specified signal regarding the intended discourse relation (Knott, 1996) than some other conjunctions. For our data, if one considers SO and BUT as compatible with author-chosen AND and allow for such matches in computation of the overall agreement rate, participant selections matched the authors' original conjunction 70% of the time. The confusion matrix for author-chosen and participant-selected conjunctions is shown in table 1.

	AND	BECAUSE	BUT	OR	SO
AND	2686	149	325	159	344
BECAUSE	280	786	176	156	156
BUT	1000	174	2798	179	180
OR	68	41	15	355	28
SO	550	127	129	298	1215
BEFORE	4	2	1	0	1
NONE	248	105	158	108	167
OTHER	8	16	10	5	9

Table 1: Confusion matrix of author conjunctions (columns) and participant responses (rows) in explicit passages

Other cases of divergence in participant selection point to contexts in which normally different conjunctions can convey the same relation. A case in point are passages containing the adverbial **otherwise** (table 2). Here, author OR received an unexpectedly high number of BECAUSE participant responses, and vice versa.

It appears that, with **otherwise**, both BECAUSE and OR can be used to express a reason. This is apparent in passage (6) below, where responses to author OR were split, with 17 participants selecting OR and 11, BECAUSE.

	AND	BECAUSE	BUT	OR	SO
AND	8	2	3	3	0
BECAUSE	31	62	11	95	4
BUT	30	7	157	2	8
OR	27	35	6	133	9
SO	2	0	4	1	3
NONE	14	6	15	18	4

Table 2: Confusion matrix for explicit passages containing **otherwise** (author conjunctions as columns, participant responses as rows)

(6) "The Ravitch camp has had about 25 fundraisers and has scheduled 20 more. Thirty others are in various stages of planning," Ms. Marcus said. "It has to be highly organized _____ otherwise it's total chaos," she added.

These two strong signals are neither noise nor disagreement nor evidence of ambiguity (as in Poesio and Artstein (2005)), but rather different, context-specific ways of conveying the same sense.

Given the number of possible responses on each trial (6 conjunctions, NONE, OTHER) and the different senses that these conjunctions are usually taken to express, our observed levels of agreement are encouraging and suggest that participants can recognize intended concurrent relations and provide meaningful responses in this task.

4.3 Responses for implicit passages

For each of the 20 adverbials in our study, we elicited responses for 25 *implicit* passages, where the original sentence contained an adverbial not preceded by a conjunction (excepting 3 implicit passages with errors in presentation). With 28 participants who saw all implicit passages, we have 13,916 analyzable data points.

To help categorize participants' behavior across adverbials, we visualize each adverbial's response profile as a stacked bar chart, as shown in figure 1 for all 20 adverbials. Every point on the x-axis represents a passage, and passages have been ordered for presentation here to highlight trends for the adverbial.[2] For each passage, bars color-coded by response (chosen conjunction) are sized according to the number of respondents who chose that response, and stacked in a consistent order: first AND (blue) at the bottom, then BECAUSE (green), then BUT (yellow), etc. The y-axis reaches 28 because

[2] In crowdsourcing the data, passages were presented in pseudo-random order (§3.3).

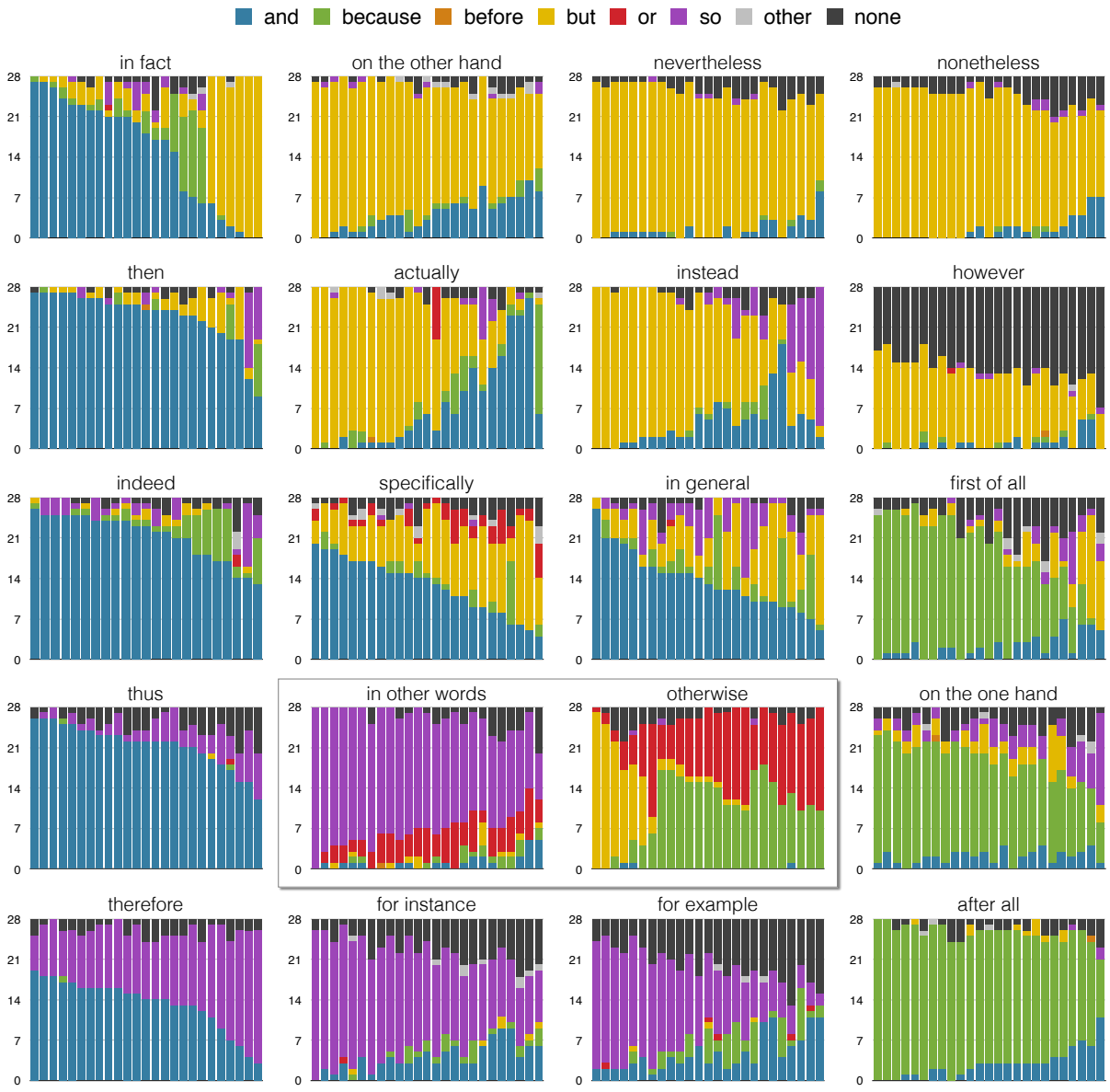

Figure 1: Data for implicit passages. Plots are arranged according to the dominant response(s).

every passage has 28 responses. We have manually arranged the plots so that patterns of dominant responses can be observed; e.g., plots with high concentrations of BUT (shown in yellow) are in the upper portion of the figure.

Comparing even just two of the plots in figure 1 leads us to several observations. Consider **otherwise** and **in other words** (highlighted in the middle of the fourth row of figure 1).

- These two adverbials have markedly different profiles of inserted conjunctions, suggesting different patterns of implied/inferred discourse relations.
- Neither response pattern is totally random; clear trends are observable in each. At the same time, neither adverbial has a single conjunction that is dominant overall. Instead, we see 2 or 3 conjunctions that are most often chosen for passages with the adverbial.
- Neither adverbial has a completely consistent distribution of responses within particular passages. The plot for **in other words** shows an overall preference on most passages for SO, but the degree of competition from BUT and OR (and even BECAUSE and AND) varies depending on the passage. For **otherwise**, some passages favor BUT whereas others are split between BECAUSE and OR responses.

We can also see that several observations in Rohde et al. (2015) regarding their 4 targeted adverbials are replicated here: The preferred conjunctions for **after all** and **in fact** are again BECAUSE

and AND/BUT/BECAUSE, respectively; likewise, **in general** has the same dominant preference for AND, although the frequency of alternative conjunctions differs. Rohde et al. reported that **instead** favored SO, but our data show SO second to BUT. This may be taken to underscore the sensitivity of these inferences to the passages in which **instead** appears.

More generally, these plots reveal striking similarities as well as striking differences. With respect to the question of whether a conjunction *can* co-occur with a discourse adverbial even when the author did not use one, the answer is yes: Participants favored the NONE option for only a few adverbials (**however**, **for instance**, **for example**), implying that the conjunctions they endorsed for other adverbials reflect connections they saw in the text and were not merely an artefact of the experiment. Furthermore, with respect to the question of how to characterize individual adverbials, figure 1 shows that neither of the uniformity outcomes listed in §4.1 hold for these data: It is not the case that all adverbials co-occur with the same preferred conjunction, nor does each have a single preferred conjunction or necessarily pattern with other adverbials from the same semantic class.

More specifically, we see that all adverbials have 1–3 frequent responses out of the 8 options. Although none the plots are overwhelmingly dominated by a single conjunction, **nevertheless** and **nonetheless** come closest with their preference for BUT. The responses BEFORE (orange) and OTHER (gray) were very rare. Some pairs of neighboring plots are highly similar, e.g., **nevertheless/ nonetheless** in the upper right, and **for instance/ for example** in the bottom center. This is reassuring as the members of each pair have intuitively similar meanings. That said, even though **actually**, **indeed**, and **in fact** would all be classified as modal stance adverbials (Aijmer and Simon-Vandenbergen, 2007), they elicit different response patterns: **actually** and **in fact** elicit AND, BUT, and BECAUSE with a smattering of SO, while **indeed** elicits AND and BECAUSE.

On the other hand, **instead** exhibits a context-specific pattern of inference: Many **instead** passages elicit a BUT response, but others elicit SO, showing that what drives the choice must be specific to the passage, not the adverbial alone. Likewise for **otherwise**: some passages elicit BUT, but most reflect an explanation, conveyed with either BECAUSE or OR, similar to responses for the explicit **otherwise** passages (§4.2).

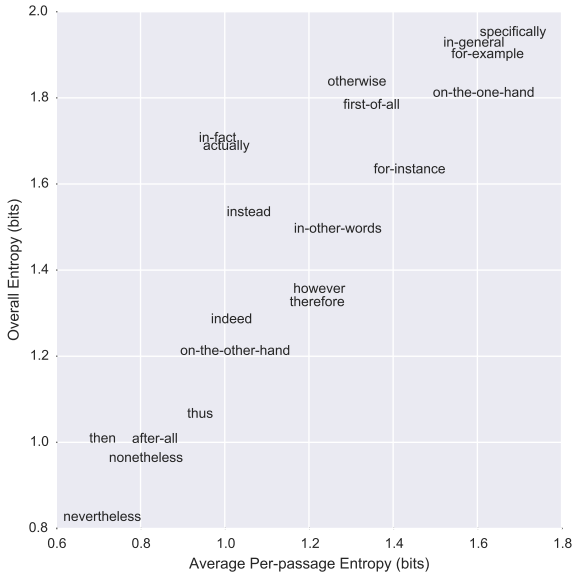

Figure 2: Each adverbial's entropy of responses for implicit passages. The x-axis is mean per-passage entropy; the y-axis is entropy of the distribution aggregating all responses over all passages for the adverbial.

Entropy. An important facet to understand in our data is the extent to which (in)consistency in responses comes from *adverbials* vs. individual *passages*. Qualitatively, we observe from figure 1 that adverbials like **therefore** have a "consistent inconsistency"—i.e., most passages produce responses split evenly between two conjunctions, so the overall response distribution looks a lot like the individual passage distributions.[3] In other cases, like **in fact**, most *passages* have a dominant response, though that response differs across passages.

The within-passage vs. overall (in)consistency can be quantified by **entropy**. Each adverbial is shown in Figure 2 with the x-axis indicating the mean entropy across of the response distribution for each passage, and the y-axis indicating the entropy of the aggregate distribution of responses across passages. The adverbials differ markedly in entropy, with extremes being **nevertheless** and **specifically**. Most adverbials have overall entropy slightly greater than mean per-passage entropy, but a few stand out as having unusually high overall

[3] One might wonder if the split for an adverbial like **therefore** reflects a split between participants who uniformly favored AND and those who uniformly favored SO. However, this does not appear to be the case: No participant chose the same conjunction for all **therefore** passages; likewise for **otherwise** (which yielded a 3-way split). Both adverbials show a cline in strength of preference for each of the dominant conjunctions. **However** was an exception with 5 participants who always responded NONE and 1 who always responded BUT.

entropy given their per-passage entropies: **in fact** and **actually** are most extreme in this regard. These are cases where individual passages are *more* consistent than the overall distribution would suggest.

Implications. Our analysis suggests that, if an annotation effort wishes to fully capture the sense relations taken to hold in the presense of discourse adverbials, it should always use multiple annotators. However, if annotation resources are limited, adverbials in the lower left of figure 2 offer the most consistency, allowing one to get away with fewer annotators. Further, if an effort wants reasonable coverage of sense relations, it should assign more annotators to adverbials whose within-passage entropy accounts in our data for most of the overall entropy (i.e., those close to the diagonal).

5 Characterization of adverbials

The notion that conjunction+adverbial combinations *could* occur has been introduced in prior work (Webber, 2013; Jiang, 2013; Rohde et al., 2015), but the range observed in our dataset is unprecedented. What does this mean for annotation schemes of discourse relations? At the very least, an annotation scheme must include the possibility that, given an adverbial, another relation, signalled by a conjunction, can also be inferred.

Our data suggests how conjunctions and adverbials combine. Although one might expect this to be limited, as §4 shows, the range of combinations far exceeds any limits imposed by *ad hoc* definitions. One might expect that the combinations can be predicted based on the semantic class of the adverbial. However, when we group the adverbials by class, we see mixed results: on the one hand, adverbials that convey exemplification (**for example**, **for instance**) pattern similarly; on the other hand, it is not the case that adverbials that convey resulting states (**thus**, **therefore**) pattern uniformly (participants endorse SO for **therefore** nearly 4 times as often as for **thus**), and our examples of modal stance adverbials (**actually**, **in fact**, **indeed**) show very different distributions.

Contrary to these hypotheses, it appears that the two parts of a conjunction+adverbial combination can contribute in different ways:

i. **Same sense:** The adverbial conveys the same lexical semantics as the conjunction (e.g., SO **thus**, in which both convey the sense that the second argument is the result of the first)

ii. **Separate sense:** The adverbial conveys distinct lexical semantics from the conjunction (e.g., SO **in other words**, in which the result sense conveyed by SO has no overlap with the restatement conveyed by **in other words**)

iii. **Parasitic sense:** The sense conveyed by the adverbial serves that conveyed by the conjunction (e.g., SO **for example**, where SO conveys a result, which is then evidenced by one or more examples)

The combinations we observe suggest that the adverbial contributes meaning, but context determines what that meaning is contributed to. When both adverbial and inferred conjunction convey the same sense, it suffices to consider the discourse relation expressed by the adverbial; otherwise, the the meaning of each must both be considered.

Finally, we turn to annotator disagreement. We define *divergent tokens* as those for which at least 8 participants chose each of two conjunctions from the set BECAUSE, BUT, OR or SO. Since AND can sometimes be taken as underspecified and hence compatible with SO and BUT, it is not included here as a fully independent competitor.

Some *divergent tokens* show annotators connecting the post-gap text to different parts of the context through different conjunctions. In the explicit passage shown in (4), 13 participants chose BECAUSE (the original author's choice) and 11 chose BUT. Closer examination reveals that different choices connect to different parts of the pre-gap context: BECAUSE links "his right foot was shaking" to the subordinate clause ("who caused everyone to laugh"), whereas BUT, like the adverbial **indeed**, links it to the statement "No, my leg's not shaking". In this case, the divergent participant choices demonstrate the disambiguating effect of the conjunction where multiple relations are possible. By removing the conjunction (which performed a different role from the adverbial), relations between the two spans are rendered ambiguous.

Other *divergent tokens* show annotators drawing different interpretations between the same spans. For the passage shown in (7), 15 participants selected BECAUSE, and 11 chose BUT (the original author's choice).

(7) There was a testy moment driving over the George Washington Bridge when the toll-taker charged him $24 for his truck and trailer, _____ after all it was New York.

Here, BUT can be taken to express a concession

with respect to the expectation that bridge tolls are usually a small amount of money (not $24), whereas BECAUSE expresses the reason why the reader should not be surprised why it's so high.

6 Conclusion and future work

We set out to gather further evidence that a semantic relationship between a sentence and its context may both be marked explicitly **and** involve inference. The extensive data we gathered through crowdsourcing judgments (20 adverbials, 50 different passages each, 28 different participants), replicate and extend earlier findings that discourse adverbials do indeed license inferred conjunctions. The patterns we have observed show that selected conjunctions are neither uniform across all 20 adverbials nor uniform within passages for a particular adverbial, but that both types of variability can be used to characterize the adverbials. In some cases, the adverbial and conjunction selected by participants share the same sense; in other cases, they are distinct (or sometimes even parasitic on the other).

Further, the diverse patterns observed for the adverbials include cases in which more than one valid connection can be inferred, each endorsed by a substantial number of participants. This resembles the *true ambiguity* of coreferential pronouns observed earlier by Poesio and Artstein (2005). Without gathering judgments from a substantial number of participants, such differences in annotation might otherwise be written off as annotator error or bias, or just a low level of inter-annotator agreement. Here, they reveal real differences in how people take a piece of text to relate to its context.

A reviewer asks if participant behavior changes over time. Because we ensured that the passages for a given adverbial appeared in a different pseudo-random order for each participant, any performance differences early or late in the token set could yield noise but not overall bias per adverbial. Token order was recorded, so future analysis is possible to test for changes in the overall rate of certain responses over time or the interactions over time between different adverbials, different participants, different conjunction-presentation orders, etc.

To extend our set of analyzed adverbials and to understand the mutual informativity between adverbials and conjunctions, another crowdsourced study with 35 new adverbials is underway, with a complementary study planned that asks participants to fill in an adverbial following a conjunction (i.e., given a conjunction, is an adverbial recoverable?). In addition, we are piloting a new response interface in which participants can select multiple conjunctions, as a means of testing whether individual participants endorse the alternative and sometimes divergent conjunctions observed across participants for a given passage.

Acknowledgments

This work has been supported in part by a grant from the Nuance Foundation. We thank Yangfeng Ji for a helpful suggestion of related work, and anonymous reviewers for their feedback.

References

Karin Aijmer and Anne-Marie Simon-Vandenbergen. 2007. *The Semantic Field of Modal Certainty: A Corpus-Based Study of English Adverbs*. Mouton de Gruyter.

Ron Artstein and Massimo Poesio. 2005. Bias decreases in proportion to the number of annotators. In James Rogers, editor, *Proceedings, 10th Conference on Formal Grammar and 9th Meeting on Mathematics of Language*, pages 139–148. CSLI Publications, Edinburgh, Scotland, UK.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596.

Maria-Josep Cuenca and Maria-Josep Marín. 2009. Co-occurrence of discourse markers in Catalan and Spanish oral narrative. *Journal of Pragmatics*, 41(5):899–914.

Bruce Fraser. 2013. Combinations of contrastive discourse markers in English. *International Review of Pragmatics*, 5:318–340.

Julia Hirschberg and Diane Litman. 1987. Now let's talk about *now*: identifying cue phrases intonationally. In *Proceedings, 25th Annual Meeting of the Association for Computational Linguistics*, pages 163–171. Stanford, California, USA.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings, 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130. Atlanta, Georgia, USA.

Xi Jiang. 2013. *Predicting the use and interpretation of implicit and explicit discourse connectives*. Ph.D. thesis, Linguistics and English Language (LEL), University of Edinburgh. MSc in English Language.

Andrew Kehler. 2002. *Coherence, Reference and the Theory of Grammar*. CSLI Publications.

Andrew Kehler and Hannah Rohde. 2013. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1–2):1–37.

Alistair Knott. 1996. *A Data-driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, Department of Artificial Intelligence, University of Edinburgh.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books Ngram Corpus. In *Proceedings, ACL 2012 System Demonstrations*, pages 169–174. Jeju Island, Korea.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter

Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182. PMID: 21163965.

Sarah Oates. 2000. Multiple discourse marker occurrence: Creating hierarchies for natural language generation. In *Proceedings, ANLP-NAACL 2000 Student Research Workshop*, pages 41–45. Seattle, Washington, USA.

Rebecca Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association of Computational Linguistics*, 2(1):311–326.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings, Workshop on Frontiers in Corpus Annotations II*, pages 76–83. Ann Arbor, Michigan.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings, 6th International Conference on Language Resources and Evaluation*, pages 2961–2968. Marrakech, Morocco.

Hannah Rohde, Anna Dickinson, Chris Clark, Annie Louis, and Bonnie Webber. 2015. Recovering discourse relations: Varying influence of discourse adverbials. In *Proceedings, First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 22–31. Lisbon, Portugal.

Evan Sandhaus. 2008. New York Times corpus: Corpus overview. LDC catalogue entry LDC2008T19.

Merel C.J. Scholman, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. 2016. A step-wise approach to discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue & Discourse*, 7(2):1–28.

Bonnie Webber. 2013. What excludes an alternative in coherence relations? In *Proceedings, 10th International Conference on Computational Semantics*, pages 276–287. Potsdam, Germany.

Bonnie Webber, Alistair Knott, and Aravind Joshi. 1999. Multiple discourse connectives in a lexicalized grammar for discourse. In *Third International Workshop on Computational Semantics*, pages 309–325. Tilburg, The Netherlands.

Deniz Zeyrek. 2014. On the distribution of contrastive-concessive discourse connectives *ama* ('but/yet') and *fakat* ('but') in written Turkish. In P. Suihkonen and L.J. Whaley, editors, *On Diversity and Complexity of Languages Spoken in Europe and North and Central Asia*.