A Discourse-Annotated Corpus of Conjoined VPs

Bonnie Webber* *University of Edinburgh Edinburgh UK bonnie.webber@ed.ac.uk

Rashmi Prasad† †University of Wisconsin-Milwaukee Milwaukee WI prasadr@uwm.edu

Alan Lee♯ **Aravind Joshi♯** ♯University of Pennsylvania Philadelphia PA [aleewk, joshi]@seas.upenn.edu

Abstract

English grammars indicate a variety of relations holding between conjoined VPs. VPs conjoined by *and* evince such senses as Result, Temporal Sequence and Concession. Although all these senses are ones associated with discourse relations, conjoined VPs have not been fully included in discourse annotation. Because of the value of discourse-annotated corpora for developing approaches to automated sense recognition, we have added their annotation to the Penn Discourse TreeBank. This paper describes how tokens were identified; how the process of span and sense annotation was modified and extended in order to keep the annotation of intra-sentential multi-clausal structures consistent with the rest of the corpus; and what the resulting corpus looks like, in terms of token frequency and common sense patterns.

1 Introduction

As frequently noted, discourse relations can hold within a sentence (i.e., *intra-sententially*) as well between larger units of text. Interest in automatically recognizing intra-sentential discourse relations (Joty et al., 2015) has recently grown e.g. to support Statistical Machine Translation (Guzmán et al., 2014) or Question Answering (Prasad and Joshi, 2008; Mannem et al., 2010). We have therefore started to expand the annotation of intra-sentential discourse relations in the Penn Discourse TreeBank (Prasad et al., 2008; Prasad et al., 2014), starting with conjoined VPs.

According to English grammar (Huddleston and Pullum, 2002), conjoined VPs can have senses other than simply **Conjunction** (and), **Disjunction** (or), and **Contrast** (but). Huddleson & Pullum note that *X and Y* may, for example, convey:

- **Consequence** (X and therefore Y), as in
  (1) Scopes was convicted and fined $100 . . . [wsj_0946]
- **Temporal Sequence** (X and then Y), as in
  (2) Tripoli says Rome kidnapped 5,000 Libyans and deported them as forced labor. [wsj_0990]
- **Concession** (despite X, Y), as in
  (3) Blacks and Hispanics currently make up 38% of the city's population and hold only 25% of the seats on the council. [wsj_1137]
- **Temporal Inclusion** (X while Y), as in
  (4) . . . the government can ensure the same flow of resources and reduce the current deficit. [wsj_1131]

Although all of these are senses usually associated with discourse relations, we have found only one corpus in which conjoined VPs have

Proceedings of LAW X – The 10th Linguistic Annotation Workshop, pages 22–31,
Berlin, Germany, August 11, 2016.

been fully treated as a locus of discourse coherence. This is a ∼53K-word corpus of home repair instructions (Subba and Di Eugenio, 2009) that was annotated according to guidelines in (Kim and Eugenio, 2006). The corpus contains ∼540 conjoined verb phrases and conjoined verbs annotated with either generic senses such as **General:Specific**, **Comparison**, **Restatement**, etc. or senses specific to the domain of instructions, such as **Criterion:act** and **Criterion:wrong-act** (depending on whether the specified action is appropriate or sub-optimal if the criterion holds). In future work, we will consider this sense annotation in more depth.

With respect to the RST Corpus (Carlson et al., 2003), its annotation guidelines[1] call for the segmentation of some but not all coordinated VPs into separate EDUs (Section 2.5.2), with only those segmented into EDUs being annotated with RST relations. With respect to the 2007 SDRT corpus, its annotation manual[2] specifies that coordinated VPs are only treated as separate discourse segments "when they either include a discourse particle or contain discourse structure within (at least one of) the coordinated constituents".

Because of the value of corpora annotated for discourse coherence for developing approaches to automated sense recognition, we decided to expand the Penn Discourse TreeBank (PDTB2) to include discourse relations associated with conjoined VPs and to package up these new annotations, along with some related annotation already in the PDTB2 (see below), for early release of a conjoined VP sub-corpus. This paper thus describes how we identified tokens to be included in the sub-corpus (Section 2); how we modified and extended the process of span and sense annotation used in the PDTB2 in order to produce annotation of intra-sentential multi-clausal structures that was consistent with the rest of the PDTB2 (Sections 3–4); and what the resulting sub-corpus looks like, in terms of inter-annotator agreement prior to adjudication, and then final token frequency and common sense patterns after adjudication was complete (Section 5).

[1] https://www.isi.edu/∼marcu/discourse/tagging-ref-manual.pdf

[2] http://timeml.org/jamesp/annotation_manual.pdf

2 Creating a Corpus of Conjoined VPs

2.1 Identifying Conjoined VPs

We took as our goal, to annotate every token in the Penn Wall Street Journal (WSJ) corpus that was analyzed as a conjoined VP in the Penn Treebank syntactic annotation of the corpus.[3] However, as Maier and colleagues have noted (2012), coordination is not reliably annotated in the PTB (or any other large treebank, for that matter). They note, in particular, that punctuation used to separate elements of a conjoined structure is annotated no differently than punctuation used for other purposes. In response, they have developed an algorithm for enhancing the annotation of punctuation used in conjoined structures.

The two-step process we used for identifying conjoined VPs did not make use of this algorithm *per se*, but something similar, focussed on conjoined VPs:

- Search the PTB parses for all sister VPs separated by a conjunction, conjunction phrase (e.g. *rather than*) or punctuation, and an optional adverbial.

- For each such pair of sisters, pre-annotate the righthand VP as **Arg2** of a potential discourse relation. If a conjunction or conjunction phrase appears between the two sister VPs, the type of the token was taken to be **Explicit** and the conjunction or conjunction phrase was labelled as the connective. If the sister VPs were separated by punctuation, the token type was taken to be **Implicit**. Later, during sense annotation (cf. Section 4), this type could be changed to **AltLex** (alternative lexicalization), if the annotators identified material in either *Arg1* or **Arg2** that made the insertion of an implicit connective seem redundant. In some cases, *Arg1* could be pre-annotated as well.

This process of pre-annotation produced false positives and false negatives, as well as true positives, all of which are informative with respect to understanding what the corpus contains.

2.1.1 False Positives (FPs)

FPs derive from two aspects of PTB analyses. The first involves ambiguous punctuation, as al-

[3] The Penn WSJ Corpus comprises the texts underlying both the Penn TreeBank (PTB) and the PDTB2.

ready noted (Maier et al., 2012), where VPs separated by comma-punctuation are not actually conjoined. The second involves tokens of *argument/adjunct cluster coordination* (Mouret, 2006; Steedman, 1989; Steedman, 2000), also called *non-constituent conjunction*, that are analyzed as conjoined VPs in the Penn TreeBank, but whose righthand conjunct lacks a verb, as in

(5) "I pay a lot to the farmer and **five times the state salary to my employees**," he says [wsj_1146]

where corresponding pairs of direct and indirect objects of *pay* have been coordinated, and

(6) She adopted 12 of assorted races, naming them the Rainbow Tribe, and driving her husband first to despair and **then to Argentina**. [wsj_1327]

where corresponding pairs of adverb and PP have been coordinated. Since they were relatively easy to recognize manually, we decided to simply exclude all such verbless VPs from the corpus.

2.1.2 False Negatives (FNs)

FNs comprise the ∼170 sequences that were analyzed in the Penn TreeBank as conjoined S-nodes with null subjects. These were discovered after completing the annotation of pre-annotated conjoined VPs, when we turned our attention to intra-sentential conjoined clauses. The tokens pre-annotated for this task were sister S-nodes separated by a conjunction, conjunction phrase (e.g. *rather than*) or punctuation, and an optional adverbial. Among the pre-annotated sister S-nodes were ones with (co-indexed) null subjects, as was the case with sentences such as the following:

(7) He joined the firm in 1963 and **bought it from the owners the next year**. [wsj_0305]

(8) The company said its directors, management and subsidiaries will remain long-term investors and **won't tender any of their shares under the offer**. [wsj_0308]

(9) The NAM embraces efforts, which both the administration and the medical profession have begun, to measure the effectiveness of medical treatments and **then to draft medical-practice guidelines**. [wsj_0314]

Since these were incorrectly analyzed according to the Penn TreeBank Guidelines (Marcus et al., 1993) and do not actually differ from the tokens already included in the corpus, we decided to include them.

On the other hand, we decided to exclude tokens containing conjoined verbs that should possibly have been analyzed as conjoined VPs, such as *exist and fight* in

```
( (S (RB Then)
   (NP-SBJ (-NONE- *) )
   (VP
     (VP (VB take)
         (NP (DT the) (VBN expected) (NN return) ))
     (CC and)
     (VP (VB subtract)
         (NP (CD one) (JJ standard) (NN deviation) )))
       (. .) ))
```

Figure 1: PTB Parse Tree for Ex. 13, showing its resemblence to the analysis of conjoined VPs

(10) The wonder is not that the resistance has failed to topple the Kabul regime , but that it continues to exist and fight at all. [wsj_2052]

We did not discover such tokens until late in the annotation process, and we lacked the resources to manually review them. It would be possible to return in the future and find and annotate them.

2.1.3 True Positives (TPs)

TPs identified through this pre-annotation process included conjoined tensed VPs (Ex. 11), conjoined adjunct VPs (Ex. 12), and conjoined imperative sentences (Exs. 13–14), which are parsed in the Penn TreeBank as conjoined VPs (Figure 1).

(11) It employs 2,700 people and **has annual revenue of about $370 million**. [wsj_0007]

(12) But many owners plan to practice frugality – crossing out the old code and **writing in the new one** until their stock runs out. [wsj_1270]

(13) Then take the expected return and **subtract one standard deviation**. [wsj_1564]

(14) Be careful boys; **use good judgment**. [wsj_0596]

2.2 Discourse Adverbials

As can be seen from the presence of *then* in Ex. 9, conjoined VPs can themselves contain discourse adverbials. As with all discourse adverbials, ones that appear in **Arg2** of a conjoined VP can link to material elsewhere in the text, as in Ex. 15

(15) Separately, the Federal Energy Regulatory Commission turned down for now a request by Northeast seeking approval of its possible purchase of PS of New Hampshire. Northeast said it *would refile its request* and **still hopes for an expedited review by the FERC** . . . [wsj_0013]

While the discourse adverbial *still* shares its **Arg2** with the conjoined VP, its *Arg1* has been taken to be the FERC turning down its request for *approval of its possible purchase of PS of New Hampshire*, which appears in the previous sentence.

Although such adverbials can link to material in previous sentences, the far more common situation (occurring in 229/230 of the VP conjuncts

in the Penn *Wall Street Journal Corpus* that contain discourse adverbials) is for such adverbials to link with the first argument *Arg1* of the conjoined VP. When they do, they serve as an explicit signal of one or more discourse relations holding between the two arguments. Among the annotated discourse adverbials from the PDTB2 found in conjoined VPs are *instead*, *still*, *then*, etc. – e.g.,

(16) He could develop the beach through a trust, but instead is trying to have his grandson become a naturalized Mexican so his family gains direct control. [wsj_0300]

(17) This year, Mr. Wathen says the firm will be able to service debt and still turn a modest profit [wsj_0305]

(18) In the engine department, several companies displayed experimental models that within a decade could provide power equal to today's engines and yet take up only half the space, ... [wsj_0956]

As such, we decided to add these tokens to the conjoined VP sub-corpus, so that one would be able to compare relations between conjoined VPs signalled with an explicit discourse adverbial with relations between them that were left implicit.

3 Labelling Arguments and their Spans

3.1 Changes to argument labelling

Early in the new annotation task, we realized that if we strictly followed the conventions used earlier in labelling arguments in the PDTB2, some span labels would be inconsistent. Here we describe what we did to overcome the problem in a way that would avoid any inconsistency.

Arguments were labelled in the PDTB2 according to the following two-part convention.

- For spans linked by an explicit discourse connective (called **explicit** relations), **Arg2** was the argument to which the connective was attached syntactically, and the other was *Arg1*. This allowed the arguments to subordinating conjunctions to be labelled consistently, independent of the order in which the arguments appeared. The same was true for coordinating conjunctions, whose argument order is always the same, and for discourse adverbials, whose *Arg1* always precedes the adverbial, even when *Arg1* is embedded in **Arg2**, as in

  (19) **A farmer** who *was kicked by his donkey* **would** nevertheless **not take revenge**.

- For spans linked by adjacency (called **implicit** discourse relations), *Arg1* was always the first (lefthand) span and **Arg2**, the second (righthand) span.

Blindly applying these same conventions *intra-sententially* produced inconsistent labelling because of (1) variability in where an explicit connectives can attach within a sentence; and (2) the ability of marked syntax to replace explicit connectives.

The first problem can be illustrated with paired connectives like *not only ... but also*. Here, both members of the pair may be present (Ex. 20), or just one or the other (Ex. 21 and Ex. 22):

(20) Japan not only outstrips the U.S. in investment flows but also outranks it in trade with most Southeast Asian countries ... [wsj_0043]

(21) The hacker was pawing over the Berkeley files but also using Berkeley and other easily accessible computers as stepping stones ... [wsj_0257]

(22) Not only did Mr. Ortega's comments come in the midst of what was intended as a showcase for the region, it came as Nicaragua is under special international scrutiny ... [wsj_0655]

A labelling convention that requires **Arg2** to be the argument to which the explicit connective attaches will choose a different argument for **Arg2** in Ex. 21 than in Ex. 22, and an arbitrary argument in the case of Ex 20, when semantically, the lefthand argument is playing the same role in all three cases, as is the righthand argument.

The second problem can be illustrated with preposed auxiliaries, which signal that a **Conditional** relation holds between the clause with the preposed auxiliary (as *antecedent*) and the other clause (as *consequent*). As with subordinating clauses, the two clauses can appear in either order:

(23) Had the contest gone a full seven games, ABC could have reaped an extra $10 million in ad sales ... [wsj_0443]

(24) ... they probably would have gotten away with it, had they not felt compelled to add Ms. Collins's signature tune, "Amazing Grace," ... [wsj_0207]

But since there is no explicit connective in either clause, if position is used to label *Arg1* and **Arg2**, the result will again be inconsistent.

A solution that addresses both these issues, while not requiring any change to existing labels in the PDTB 2.0 is the following:

- The arguments to inter-sentential discourse relations remain labelled by their *position*: *Arg1* is first (lefthand) argument and **Arg2**, the second (righthand) argument.

- With intra-sentential coordinating structures, the arguments are also labelled by their *position*: *Arg1* is first argument and **Arg2**, the second one.

- With intra-sentential subordinating structures, *Arg1* and **Arg2** are determined syntactically. The subordinate structure is always labelled **Arg2**, and the structure to which it is subordinate is labelled *Arg1*.

3.2 Changes to span-labelling

In PDTB2 annotation, the arguments to relations are text spans. But the text span(s) that make up an argument are required to subsume *at least one full clause*, including parts of the clause that might not be *relevant* to the relation. While this continues to be the guideline for annotating non-coordinating constructions, for coordinating constructions, the guideline has been changed such that annotators are asked to annotate just the conjuncts, which in the case of conjoined VPs is *not* a whole clause. Thus, in Ex. 7, *Arg1* subsumes only *joined the firm in 1963*, and not the subject *he*. The same goes for Ex. 11.

A second change involves *relevance*: Annotators were told that material that contributes semantically to both arguments of a conjoined VP should be omitted, so that it is not taken to be specific to one argument or the other. The result is that spans in the corpus may not completely match the spans of VPs in the Penn TreeBank. For example, in

(25) UAL ...reversed course and plummeted in off-exchange trading after the 5:00 p.m. EDT announcement. [wsj_1305]

the PTB takes *reversed course* as being conjoined with *plummeted in off-exchange trading after the 5:00 p.m. EDT announcement*, even though both reversing course and plummeting happened *in off-exchange trading after the 5:00 p.m. EDT announcement*. Recognizing this, the annotators changed the second conjunct to *plummeted*.

Annotators were also told that the spans of both arguments should be parallel — both bare infinitives, or to-infinitives, or tensed clauses, etc. So in Ex. 9, since **Arg2** is the to-infinitive *then to draft medical-practice guidelines*, selected as *Arg1* would be the to-infinitive *to measure the effectiveness of medical treatments*.

Also common among conjoined VPs are *attribution phrases* such as *said* and *added* in Ex. 26 and *declare* in Ex. 27. When annotating implicit relations on conjoined VPs, annotators were told to retain only those attributions that contribute to the semantics of the relation (as in Ex. 27, where the **Purpose** of declaring something a pesticide is so that it can be pulled from the marketplace). In Ex. 26, neither *said* nor *added* contribute to the **Concession** relation that is taken to hold, so annotators omitted them from the spans of *Arg1* and **Arg2**.

(26) The company, based in San Francisco, said *it had to shut down a crude-oil pipeline in the Bay area to check for leaks* but added that **its refinery in nearby Richmond, Calif., was undamaged**. [wsj_1884]

(27) Give the EPA more flexibility to *declare a pesticide an imminent hazard* and **pull it from the marketplace**. [wsj_0964]

The final thing to say here about attribution is that where an annotator takes the same relation to hold between attribution phrases as between content of attribution, we ask that the relation be annotated between the latter, indicating the minimal spans that give rise to the particular relational sense.

4 Labelling Relation Senses

4.1 Changes to the Relation Hierarchy

We have extended and simplified the PDTB2 relation hierarchy, producing a new PDTB3 relation hierarchy (Figure 2). Some of the changes (such as restricting Level-3 relations to differences in directionality, eliminating rare and/or difficult-to-annotate senses, and replacing separate senses with features that can be added to a given sense) are meant to simplify annotation (Section 4.1.1). Other changes are additions to the relation hierarchy motivated by the intra-sentential relations we have been annotating, including ones associated with conjoined VPs (Section 4.1.2).

4.1.1 Simplifying the relation hierarchy

Although the hierarchy retains the same four Level-1 relations, relations at Level-3 now only encode *directionality* and so only appear with asymmetric Level-2 relations.[4] Those Level-3 relations in the PDTB2 that did not convey directionality were either moved to Level-2 — **Substitution** (renamed from the PDTB2 **Chosen Alternative**) and **Equivalence** — or eliminated due to their rarity or the difficulty they posed for annotators — in particular, those under the Level-2 relations of **Contrast**, **Condition** and **Alternative** (now renamed **Disjunction**).

With respect to directionality, annotating additional intra-sententential discourse relations has called attention to asymmetric Level-2 relations

[4] A sense relation $\Re$ is *symmetric* iff $\Re$(*Arg1*, **Arg2**) and $\Re$(**Arg2**, *Arg1*) are semantically equivalent. If a relation is not symmetric, it is *asymmetric*.

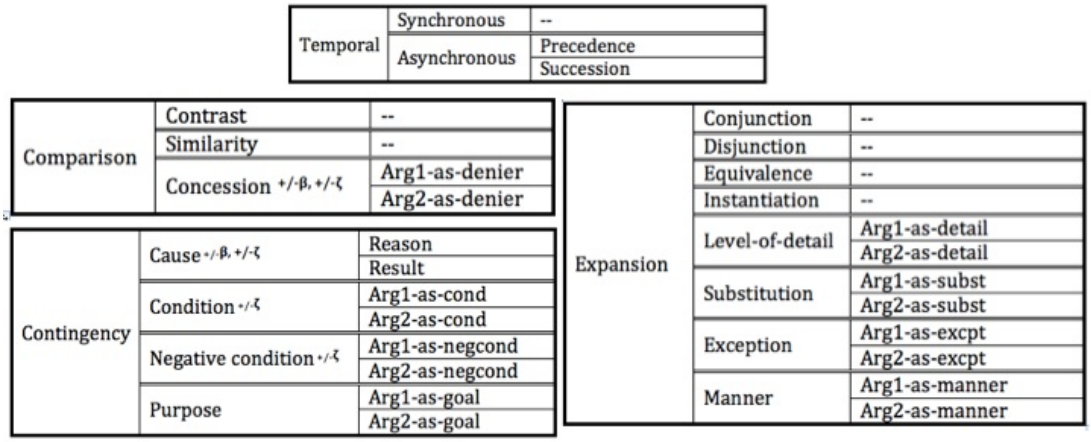

Figure 2: PDTB3 Relation Hierarchy. Only asymmetric relations are specified further at Level-3, to capture the directionality of the arguments. Superscript symbols on Level-2 senses indicate features for implicit beliefs (+/-β) and speech-acts (+/-ζ) that may or may not be associated with one of the defined arguments of the relation. Features are shown on the relation in the table here only for clarity, but should not be seen as a property of the relation, rather of the arguments.

whose arguments have been found to occur in either order (rather than the single order assumed in the PDTB2). In particular, the argument conveying the condition in **Condition** relations can be either **Arg2** (as was the case throughout the PDTB2) or *Arg1* as in Ex. 28, while the argument conveying the "chosen alternative" (now called "substitute") in **Substitution** relations can be either **Arg2** (as was the case throughout the PDTB2) or *Arg1*, as in Ex. 29. In the case of the rare relation called **Exception**, it was not previously noticed that in some of the tokens so annotated, the exception appeared in **Arg2**, while in the rest, the exception appeared in *Arg1*. The difference is now supported with a distinct Level-3 type in each direction (Exs. 30–31).

(28) **Arg1-as-cond:** *Call Jim Wright's office in downtown Fort Worth, Texas, these days* and **the receptionist still answers the phone, "Speaker Wright's office.** [wsj_0909]

(29) **Arg1-as-subst:** "The primary purpose of a railing is *to contain a vehicle* and **not to provide a scenic view**," [wsj_0102]

(30) **Arg1-as-excpt:** *Twenty-five years ago the poet Richard Wilbur modernized this 17th-century comedy merely by avoiding "the zounds sort of thing," as he wrote in his introduction*. Otherwise, **the scene remained Celimene's house in 1666**. [wsj_1936]

(31) **Arg2-as-excpt:** *Boston Co. officials declined to comment on Moodys action on the units financial performance this year* except **to deny a published report that outside accountants had discovered evidence of significant accounting errors in the first three quarters results**. [wsj_1103]

Level-2 pragmatic relations have been removed from the PDTB2 and replaced with features that can be attached to a relation token to indicate an inference of *implicit* belief (epistemic knowledge) or of a *speech act* associated with arguments, rather than with the relation itself. Figure 2 shows the relations for which these features have so far been found to be warranted, based on the empirical evidence found during annotation. Ex. 32 shows an implicit **Cause.Result** relation but one where the result **Arg2** argument is the (speaker's/writer's) *belief* that the deadline could be extended. **Arg2** is therefore annotated with a +belief feature because the belief is implicit. Similarly, Ex. 33 shows a **Concession.Arg2-as-denier** relation, but what's being denied (or cancelled) is the speech act associated with **Arg2**, and this is annotated as a feature on **Arg2** because it is implicit.

(32) That deadline *has been extended once* and Implicit=so **could be extended again**. [wsj_2032]

(33) He spends his days *sketching passers-by*, or **trying to**. [wsj_0039]

Also simplifying the PDTB2 hierarchy is removal of the **List** relation, which does not appear semantically distinguishable from **Conjunction**. And the names of two asymmetric PDTB2 relations have been changed to bring out commonalities. In particular, **Restatement** has been renamed **Level-of-detail**, with its **Specification** and **Generalization** subtypes in the PDTB2 now just taken to be directional variants renamed **Arg2-as-detail**

and **Arg1-as-detail**, respectively; and the subtypes of **Concession**, opaquely called **Contra-expectation** and **Expectation**, have been renamed to reflect simply a difference in directionality: **Arg1-as-denier** and **Arg2-as-denier**.

4.1.2 Augmenting the relation hierarchy

Additional senses found to be needed for annotating conjoined VPs include **Manner** under **Expansion** (both Level-3 directions), and **Negative_Condition** and **Purpose** under **Contingency** (with both Level-3 directions for each). The new symmetric Level-2 relation of **Similarity** (under **Comparison**) was added because of its obvious omission from the PDTB2 as the complement of the symmetric relation **Contrast**.

Definitions and examples for these new relations are given in Table 1.

Note that the entire PDTB2 is being mapped to senses in the revised relation hierarchy, not just the conjoined VP sub-corpus. Most often, the mapping is simply 1:1. Where the mapping is 1:N or M:N, manual review has been required, with further adjudication to ensure both agreement and consistency. When the PDTB3 is released to the public in September 2017, we will record the frequency with which each PDTB2 sense has been replaced by a specific PDTB3 sense.

4.2 Sense labelling of conjoined VP tokens

The VPs presented to annotators were conjoined either lexically or by punctuation. The annotators were given guidelines for assigning sense relations that depended on the particular configuration involved — specifically:

1. An explicit conjunction can have a single sense, which can be **Conjunction** (Ex. 34), or something else (Ex. 35-36).
   (34) The concept may be simple: Take a bunch of loans, *tie them up in one neat package*, and **sell pieces of the package to investors**. (**Expansion.Conjunction**) [wsj_1635]
   (35) These active suspension systems *electronically sense road conditions* and **adjust a car's ride** (**Contingency.Purpose.Arg2-as-goal**) [wsj_0956]
   (36) Stocks *closed higher in Hong Kong, Manila, Singapore, Sydney and Wellington*, but **were lower in Seoul**. (**Comparison.Contrast**) [wsj 0231]

2. The arguments to an explicit conjunction can also be linked by an additional relation, conveyed implicitly (Ex. 37-38) or by an explicit discourse adverbial. (Such adverbials were taken to have been already annotated in PDTB2.) To indicate an additional implicit relation, annotators created a new annotation token for the same two conjuncts, inserted an appropriate implicit connective and labeled it with the sense(s) they inferred. Argument spans of the explicit and the implicit relation were *not* required to be the same, so annotators could adjust the spans of the new token if needed.
   (37) We've got to *get out of the Detroit mentality* and Implicit=instead **be part of the world mentality**," declares Charles M. Jordan, GM's vice president for design ... (**Expansion.Conjunction, Expansion.Substitution.Arg2-as-subst**) [wsj_0956]
   (38) ... Exxon Corp. *built the plant* but Implicit=then closed it in 1985. (**Comparison.Concession.Arg2-as-denier, Temporal.Asynchronous.Precedence**) [wsj_1748]

3. If inserting an implicit connective was perceived as redundant, appropriate material in **Arg2** could be annotated as *AltLex* (Ex. 39), as done elsewhere in the PDTB2 (Prasad et al., 2010).
   (39) His policies *went beyond his control* **and resulted . . . in riots and disturbances**. (**Expansion.Conjunction, Contingency.Cause.Result**) [wsj_0290]

The second guideline above points to a new feature of our discourse annotation: While multiple relations were annotated in the PDTB2 as holding between identical or overlapping argument spans, all were associated with either multiple explicit connectives or multiple inferred relations. What is new in the annotation of conjoined VPs is the possibility of an explicit relation co-occurring with ones that are inferred (implicit relations). We expect to identify more of these in other syntactic contexts.

5 Corpus Characteristics

For annotation, the pre-annotated tokens were divided into 25 batches. After a batch was annotated by two annotators, inter-annotator agreement was calculated (see below), and then adjudication was carried out, for the annotators and authors to reach agreement. Annotated tokens of discourse adverbials in **Arg2** of the conjoined VPs were imported from the PDTB2 (Section 2.2), with sense labels automatically updated to reflect the revised relation hierarchy (Section 4) if there was a 1:1

Similarity: One or more similarities between *Arg1* and **Arg2** are highlighted with respect to what each argument predicates as a whole or to some entities it mentions.	. . ., *the Straits Times index is up 24% this year*, so investors who bailed out generally did so profitably. Similarly, **Kuala Lumpur's composite index yesterday ended 27.5% above its 1988 close**. [wsj_2230]
Negative Condition: One argument describes a situation presented as unrealized (the antecedent or condition), which if it doesn't occur, would lead to the situation described by the other argument (the consequent).	**Arg1-as-negcond:** In Singapore, a new law requires smokers to *put out their cigarettes before entering restaurants, department stores and sports centers* or **face a $250 fine**. [wsj_0037]
Purpose: One argument presents an action that an agent undertakes with the purpose (intention) of achieving the goal conveyed by the other argument.	**Arg1-as-goal:** She ordered *the foyer done in a different plaid planting*, and Implicit=for that purpose **made the landscape architects study a book on tartans**. [wsj_0984]
Manner: The situation described by one argument presents *how* (i.e., the manner in which) the situation described by other argument has happened or is done.	**Arg1-as-manner:** He argued that program-trading by roughly 15 big institutions is *pushing around the markets* and Implicit=thereby **scaring individual investors**. [wsj_0987]

Table 1: New relations in PDTB3

mapping between a discontinued PDTB2 sense label and its corresponding new PDTB3 label. If there wasn't a 1:1 mapping, the sense label was left empty and the annotation tool would flag the token as requiring a new sense label. The span annotations of each token were also modified to accord with the new span guidelines (Section 3.2).

The corpus comprises 3372 conjoined VPs annotated with a single sense and 1261 annotated with multiple senses. Each discourse relation is recorded as an *annotation token*, with multi-sense conjoined VPs recorded as either two linked annotation tokens (each with one or more senses) or as a single annotation token with multiple senses. In total, the corpus comprises 5894 annotation tokens.

Prior to adjudication, inter-annotator agreement (IAA) on sense annotation (full agreement on one or more senses) was 74%. Partial agreement on at least one sense was 74.3%. IAA on both senses and argument spans was 69.8%. Partial IAA on at least one sense and span was 70.1%. Of the 658 sense disagreements, the most common involved Contrast and Concession.Arg2-as-denier (127/658 =19.3%). We did not consider as disagreements, cases where only one annotator reported an additional inferred sense: On review, the other annotator acknowledged simply not noticing it.

5.1 Single-sense Conjoined VPs

Of the 3372 single-sense relations in the corpus, 2962 are lexically-conjoined VPs (2933 Explicit conjunctions and 29 Explicit adverbials) and 410 are punctuation-conjoined VPs.

Among these single-sense relations, **Expansion.Conjunction** is the most common sense, but other senses occur fairly often as well, as shown in Table 2 for Explicit conjunctions and Table 3 for punctuation-conjoined relations.

The most common single-sense Explicit connectives are *and*, *but* and *or*. While explicit *and* has **Expansion.Conjunction** as its most common sense, its senses still show the kind of variability noted in Section 1, as shown in Table 4. The most common Implicit connectives are *and*, *then* and *or*. Also relatively frequent is the use of *not* as an AltLex with the sense of **Substitution**, as in Ex. 29.

5.2 Multi-sense Conjoined VPs

As noted in Section 4.2, more than one sense may hold between the arguments of a conjoined VP, either through inference or through the presence of an explicit discourse adverbial in **Arg2**.

The corpus contains 214 Explicit adverbials linked to an Explicit conjunction, sharing their arguments. Table 5 shows the distribution of these Explicit conjunction+adverbial pairs and Table 6 their associated sense pairings.

Annotators also inferred multiple senses on conjoined VPs in the absence of an explicit adverbial. In most cases, such inferences are annotated either as a separate Implicit or AltLex tokens linked to a token containing the Explicit conjunction, while multiple senses could also be recorded on a single annotation token. Annotators inferred 53 different Implicit connectives or AltLex text spans in these cases, the most common being *then*, *therefore*/*as a result*, *thereby* and *instead*. There are 1047 such multi-sense conjoined VPs in the corpus, with the main sense pairings shown in Table 7.

In total, the corpus contains 1261 multi-sense conjoined VPs. In most cases, these multi-sense relations are annotated via the linking of two or

Sense	Frequency
Expansion.Conjunction	2113
Comparison.Concession.Arg2-as-denier	320
Expansion.Disjunction	219
Contingency.Purpose.Arg2-as-goal	106
Comparison.Contrast	93
OTHER	82
TOTAL	**2933**

Table 2: Sense distribution of single-sense lexically-conjoined conjunctions

Sense	Frequency
Expansion.Conjunction	290
Temporal.Asynchronous.Precedence	51
Expansion.Substitution.Arg2-as-subst	14
Expansion.Disjunction	14
Expansion.Equivalence	12
OTHER	29
TOTAL	**410**

Table 3: Sense distribution of single-sense punctuation-conjoined VPs

more tokens, with these links explicitly marked in the annotation files.

6 Future Work

We plan to release the corpus in two forms, for the Linguistic Annotation Workshop in August 2016. For researchers with access to the Penn TreeBank, the corpus will be available as stand-off annotation. For those lacking access to the Penn TreeBank, we will provide a limited version of the corpus containing just those sentences that contain conjoined VPs, with annotation of their spans and senses. While we will be continuing to further enrich the PDTB, the goal of this early release of a corpus of conjoined VPs is to encourage research targetted at shallow discourse parsing of these constructions, given how common they are and how useful recognition of the relations expressed in them might prove.

Acknowledgments

This work has been supported by the National Science Foundation (NSF) under grants RI 1422186 and RI 1421067.

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure The-

Sense	Frequency
Expansion.Conjunction	2032
Contingency.Purpose.Arg2-as-goal	106
Expansion.Manner.Arg2-as-manner	18
Expansion.Substitution.Arg1-as-subst	5
OTHER (4)	8
TOTAL	**2169**

Table 4: Common senses of Explicit *and*

Connective	Frequency
and + then	71
and + thus	18
and + also	15
and + later	11
and + therefore	10
OTHER (e.g. but+also, but+instead)	89
TOTAL	**214**

Table 5: Distribution of Explicit conjunction and adverbial pairs

Sense	Frequency
Conjunction + Precedence	94
Conjunction + Result	44
Conjunction + Conjunction	19
Conjunction + Arg2-as-denier	14
Arg2-as-denier + Precedence	9
OTHER	34
TOTAL	**214**

Table 6: Distribution of sense pairs associated with Explicit conjunction and adverbial pairs

Sense	Frequency
Conjunction + Result	402
Conjunction + Precedence	378
Conjunction + Arg2-as-subst	51
Conjunction + Arg2-as-detail	44
Result + Arg1-as-manner	41
OTHER	131
TOTAL	**1047**

Table 7: Distribution of sense pairs inferred on an explicit conjunction

ory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*, pages 85–112. Kluwer.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics)*, pages 687–698, Baltimore, Maryland, June.

Rodney Huddleston and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge UK.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41:385–435.

Su Nam Kim and Barbara Di Eugenio. 2006. Coding scheme for instructional corpus: Identifying segments, relations and minimal units. Technical report, University of Illinois at Chicago, March.

Wolfgang Maier, Sandra Kübler, Erhard Hinrichs, and Julia Kriwanek. 2012. Annotating coordination in the Penn Treebank. In *Proceedings, 6th Linguistic Annotation Workshop*, pages 166–174, Jeju, Republic of Korea.

Prashanth Mannem, Rashmi Prasad, and Aravind Joshi. 2010. Question generation from paragraphs at UPenn: QGSTEC system description. In *Proceedings, Third Workshop on Question Generation*.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn TreeBank. *Computational Linguistics*, 19(2):313–330.

Francois Mouret. 2006. A phrase structure approach to argument cluster coordination. In Stephan Müller, editor, *Proceedings, 13th International Conference on Head-Driven Phrase Structure Grammar*, pages 247–267.

Rashmi Prasad and Aravind Joshi. 2008. A discourse-based approach to generating why-questions from texts. In *Proceedings, Workshop on Question Generation Shared Task and Evaluation Challenge*.

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.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Realization of discourse relations by other means: Alternative lexicalizations. In *Proceedings, International Conf. on Computational Linguistics (COLING)*.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse TreeBank, comparable corpora and complementary annotation. *Computational Linguistics*, 40(4):921–950.

Mark Steedman. 1989. Constituency and coordination in combinatory grammar. In M. Baltin and A. Kroch, editors, *Alternative Conceptions of Phrase Structure*, pages 201–231. University of Chicago Press.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press.

Rajen Subba and Barbara Di Eugenio. 2009. An effective Discourse Parser that uses Rich Linguistic Information. In *Proceedings, NAACL-HLT2009*, pages 566–574.