Extracting PDTB Discourse Relations from Student Essays

Kate Forbes-Riley and **Fan Zhang** and **Diane Litman**
University of Pittsburgh
Pittsburgh, PA 15260 USA
katherineforbesriley@gmail.com, faz23@pitt.edu
dlitman@pitt.edu

Abstract

We investigate the manual and automatic annotation of PDTB discourse relations in student essays, a novel domain that is not only learning-based and argumentative, but also noisy with surface errors and deeper coherency issues. We discuss methodological complexities it poses for the task. We present descriptive statistics and compare relation distributions in related corpora. We compare automatic discourse parsing performance to prior work.

1 Introduction

The Penn Discourse Treebank (PDTB) framework (Prasad et al., 2014) has been used to add discourse relation annotation to numerous corpora, including the Wall Street Journal corpus. It differs from other approaches because of its focus on the lexical grounding of discourse relations, such that all discourse relations either are or can be instantiated by a discourse connective (e.g., *however, in other words*). This linkage between lexicon and discourse relation has been shown to yield reliable human annotation across languages (Al-saif and Markert, 2011; Zhou and Xue, 2015; Zeyrek et al., 2013; Sharma et al., 2013; Polkov et al., 2013; Danlos et al., 2012) and as a result has facilitated the increased use of discourse relations in language technology and psycholinguistics research (e.g. (Ghosh et al., 2012; Patterson and Kehler, 2003; Torabi Asr and Demberg, 2013)). Researchers are also working towards automating PDTB annotation, although performance to date is still low, with F1 scores near 30% under the strictest evaluation terms (e.g., (Lin et al., 2014; Xue et al., 2015; Ji and Eisenstein, 2014)).

The purpose of the present study is to investigate the manual and automatic annotation of PDTB relations in a corpus of student essays. This corpus differs markedly from all prior ones to which the PDTB framework has been applied. First, it is both argumentative and learning-based: students are learning about argumentative writing through the essay-writing process. Second it is noisy, displaying not only spelling and grammar errors but also deeper problems of referential and relational coherency. We hypothesized that these differences would shed light on unclear aspects of the PDTB framework, while also challenging an automatic discourse parser. However, if despite their inherent noise, learning-based datasets could be shown able to be reliably annotated for discourse relations, then they could provide language technology and psycholinguistics research a wealth of new applications. For example, interactions between students' discourse relation use and their quality and quantity of learning and affective states could be investigated (c.f. (Litman and Forbes-Riley, 2014)), as could the use of discourse relations for improving automated essay graders and writing tutors (c.f. (Zhang et al., 2016)).

In this paper we discuss methodological complexities posed by applying the PDTB framework to noisy, learning-based, and argumentative data, including a heightened ambiguity between EntRel, Expansion, and Contingency relations. We present descriptive statistics showing how the relation distributions compare to both the PDTB (Prasad et al., 2014) and BioDRB corpus (Prasad et al., 2011), whose texts possess argumentative structure without being noisy or learning-based. Some of these results suggest targets for future learning research. For example, the essays contain 12% fewer explicit connectives, contributing not only to the lowered coherency but also reflecting inexperience with connective use. We then investigate the performance of the Lin et al. (2014) PDTB-trained parser, and find that relaxing the minimal

Proceedings of the SIGDIAL 2016 Conference, pages 117–127,
Los Angeles, USA, 13-15 September 2016.

argument constraint and predicting only Level-1 senses tempers the negative impact of the noise; the parser yields an end-to-end F1 score of 31% under strictest evaluation terms, similar to other corpora and parsers (Xue et al., 2015). Like this prior work, performance is highest on the first steps of connective identification and argument match. Patterns of errors in the remaining steps indicate training on domain-specific data could help, and also that parser and human find the same relations ambiguous. Overall our results suggest that despite the inherent noise, learning-based datasets can be reliably annotated for discourse relations.

2 Student Essay Data

Most prior PDTB applications have focused on the published news domain, although the Turkish DB (Zeyrek et al., 2013) also used published novels, while the BioDRB (Prasad et al., 2011) used published biomedical research articles.

The present study uses first and second drafts of 47 AP English high school student essays (94 essays, 4271 sentences, 75900 words) (Zhang and Litman, 2015). The first drafts were written after students read the first five cantos of Dante's Inferno, and required explaining why a contemporary should be sent to each of the first six sections of Dante's hell. The second drafts were revisions by the original writers after they received and generated peer feedback as part of the course.

The essays differ markedly from news articles both in possessing an argumentative structure and being learning-based, with the goal that by the second draft they consist of an introduction, intermediate paragraphs developing the reasoning for each contemporary's placement in hell, and a conclusion. Although such over-arching rhetorical structure is deliberately ignored in the PDTB, Prasad et al. (2011) concluded that it still impacts relation distribution after applying the framework to the BioDRB, whose biomedical articles are also argumentative and segmented into introduction, method, results and discussion (IMRAD).

The student essays further differ from all prior PDTB applications in that they are noisy, containing not only grammar and spelling errors but also deeper problems of referential and relational coherency. The noise often does not improve between first and second drafts. A.1-A.4 in the appendix provide essay excerpts illustrating noise variations. As shown, not only are spelling and grammar errors common, but a comparison of A.1 and A.2 (beginning an essay) and A.3 and A.4 (mid-essay) illustrate how the lack or misuse of cohesive devices, along with weakness in or deviation from argumentative structure, creates semantic ambiguity and reduces referential and relational coherence (Sanders and Maat, 2006).

3 Manually Annotating PDTB Relations

Central tenets of the PDTB framework are its focus on the lexical grounding of discourse relations and its neutrality with respect to discourse structure beyond seeking two abstract object arguments for all relations (Prasad et al., 2014). Five relation *types* are annotated: EXPLICIT, IMPLICIT, ALTLEX, ENTREL, NOREL. Four Level-1 *senses* are annotated: COMPARISON, CONTINGENCY, EXPANSION, TEMPORAL. Level-2 and -3 senses are also annotated, along with the relation's two minimal argument spans, and when applicable, the explicit or inserted implicit connective that signals it, as well as its attribution (i.e., speaker).

Annotated essay examples for each relation type and Level-1 sense are in Appendix A.5 and below. In each, the lexical grounding of one relation is underlined (it may be implicit, explicit or alternatively lexicalized), its syntactically bound argument (ARG2) is **bolded**, its non-structural argument (ARG1) is *italicized*, and its type and sense (where applicable) are in parenthesis.

3.1 Method

Prior applications of the PDTB framework have adopted its central tenets and most of its annotation conventions while adapting others to suit language and domain (Prasad et al., 2011; Alsaif and Markert, 2011; Zhou and Xue, 2015; Zeyrek et al., 2013; Sharma et al., 2013; Polkov et al., 2013; Danlos et al., 2012). Prasad et al. (2014) provide a comparative discussion of this prior work. Following this work we too retained PDTB's central tenets and adhered to most of its annotation conventions but modified some to fit our domain, increase reliability, and reduce cost:

a) As in the Hindi DRB (Sharma et al., 2013), our workflow proceeded in one pass through each essay, with each relation annotated for type, argument span, and sense before moving on.

b) As in the BioDRB (Prasad et al., 2011), we did not label attribution, as apart from Dante quotes the student was nearly always the speaker.

c) We only labeled Level-1 senses because our noisy conditions often made finer distinctions ambiguous. We did not adopt the BIoDRB's new argument-oriented senses because it is unclear how they all map to PDTB senses[1].

d) The PDTB's STRUCTURAL ADJACENCY CONSTRAINT requires Implicits to take arguments from adjacent units. This exacerbated annotation difficulty in our noisy conditions by favoring weak relations often ambiguous between Implicit, EntRel, or NoRel over stronger non-adjacent ones. Thus as in the BioDRB we permitted Implicit non-structural arguments in non-adjacent within-paragraph units, even though the automatic parser would not. This case is illustrated in Example 1.

(1) *In the place of the hoarders houses Mary who took in too much and did not relinquish these treasures.* Dante states in Canto seven line forty-seven "Are clerks – yea, popes and cardinals, in whom covetousness hath made its masterpiece". So **Although not understanding why Gods men are housed in this circle she is sentenced to this as she is also a strong believer in God and his ways.** (Implicit/Contingency)

e) The PDTB's MINIMAL ARGUMENT CONSTRAINT requires labeling only the minimal necessary argument text. Because our noisy conditions often made boundaries ambiguous, we did not strictly enforce this. In hard cases a larger unit was labeled with the expectation that minimality could be pursued on a subsequent pass. This case is also illustrated above.

f) Often in the essays relations hold between ungrammatical units, including sentences concatenated without punctuation or syntactically incomplete ones, as illustrated in Example 2. Due to their frequency, we decided to annotate them even if the automatic parser would not.

(2) *The first layer of hell is the vestibule in the entrance of hell* **this is a large open gate symbolizing that is easy to get into.** (EntRel)

The annotation was performed using the PDTB tool from the website. The lists of connectives from the PDTB manual were used to help identify and insert implicit connectives. Although these lists are productive, only rarely was a new connective inserted, because the conditions regarding connective classification are still unclear. [2]

[1]Prasad et al. (2011) state that Continuation and Background map to Expansion and are reformulations of EntRel.

	Exp	**Imp**	**AltL**	**EntR**	**NoR**	**n/a**
Exp	**73**					8
Imp		**56**	1	13		
AltL			**1**			
EntR		2		**10**		
NoR					**0**	
n/a	2	1		1		n/a

Table 1: Relation Types in Interannotator Agreement Study (SE in rows; PDTB in cols)

	Comp	**Cont**	**Expn**	**Temp**
Comp	**28**	1		
Cont		**24**	2	
Expn	6	4	**42**	1
Temp		1		**21**

Table 2: Senses for Agreed Types in Interannotator Agreement Study (SE in rows; PDTB in cols)

3.2 Interannotator Reliability Study

The annotator used here was one of the early developers of the D-LTAG environment that engendered the PDTB framework (Forbes-Riley et al., 2006; Miltsakaki et al., 2003; Forbes et al., 2002), and was thus viewed as an expert. To verify this presumption an inter-annotator agreement study was performed. Four WSJ articles[3] were annotated for the five relation types and the four Level-1 senses and compared with the gold-standard annotations. The student essay (SE) annotator produced 163 relations while the PDTB produced 160, yielding a total of 168 unique relations. 140 agreed for relation type, meaning the type label matched and the argument spans were overlapping, i.e. an exact or partial match. Table 1 shows the type labels across the SE (rows) and PDTB (columns), with the final column/row ("n/a") representing relations identified by only one. Table 2 shows the senses for the 130 agreed types, excluding the 10 EntRels, which take no sense label. For type, agreement is 140/168, or 83%, and for sense it is 115/130, or 89%, with a Kappa of .84.

This level of agreement is on par with prior PDTB annotations. For example in the BioDRB agreement for Explicit and AltLex is 82% (Implicit agreement is not reported), and Kappa for

[2]E.g. two common clause subordinators, "by" and "in order to," are annotated in the BioDRB but not the PDTB.

[3]wsj1000, wsj2303, wsj2308, and wsj2314

Type	Count (%)	Comp	Cont	Exp	Tmp	Comp/ Cont	Comp/ Exp	Comp/ Tmp	Cont/ Exp	Cont/ Tmp	Exp/ Tmp
Exp	1657 (33%)	315	626	474	192	1	1	7	1	33	7
Imp	2495 (49%)	186	739	1492	18	2	8	4	36	4	6
AltL	103 (2%)	1	49	51	0	0	0	0	0	0	2
EntR	844 (17%)	- -	- -	- -	- -	- -	- -	- -	- -	- -	- -
NoR	0	-	-	-	-	-	-	-	-	-	-
All	5099	502	1414	2017	210	3	9	11	37	37	15
Exp/Imp Senses	4262	12%	34%	48%	6%						

Table 3: Relation Type and Sense Distribution in Student Essays

	Exp	Imp	AltL	EntR	NoR	Comp	Cont	Expn	Tmp
PDTB	45%	40%	2%	13%	0.6%	23%	22%	42%	13%
BioDRB	45%	51%	3%	0%	0.5%	11%	20%	?	17%

Table 4: Comparison Percentages of Types and Senses in PDTB Corpus and BioDRB Corpus

the 31 BioDRB senses is .71 for Explicit and AltLex and .63 for Implicit (Prasad et al., 2011). In the PDTB agreement was only reported for argument spans because some types were developed as the annotation went along. Agreement for partial match arguments is 94.5% and 92.6% for Explicits and Implicits, respectively (Miltsakaki et al., 2004; Prasad et al., 2008), while sense agreement is 94% for Level-1, falling to 84% and 80% for Level-2 and -3, respectively (Prasad et al., 2008).

3.3 Manual Annotation Results

Table 3 shows the distributions of manually annotated discourse relations in the essays. Type counts in the second column are broken down into senses across the remaining columns. As shown, Explicit, Implicit and AltLex can have multiple senses simultaneously. Table 4 compares relation distributions in the PDTB and BioDRB corpora.

Considering first relation type, there are 12% fewer Explicits in the essays than in the PDTB and BioDRB, both of which report 45%. That high school students are less likely to provide explicit markers of their intended discourse relations not only contributes to lowered coherency but also reflects their inexperience with the use of these cohesive devices, and points to an area for future learning-based language technology research. The type counts are recovered across Implicits and EntRels, with the essays containing 49% and 17%, while the PDTB contains only 40% and 13%, respectively. In the BioDRB, the addition of new senses inflated the percentage of Implicits (51%) by removing EntRels completely. AltLex appears only rarely at 2-3% across all three corpora; however, these are underannotated in the framework, i.e. only when inserting a connective creates semantic redundancy (Prasad et al., 2014). NoRels are even more rare, occurring in the PDTB and BioDRB at rates of 0.5-0.6%, and not at all in the essays. A major reason was our loosening of the structural adjacency requirement (Section 3.1)[4]; most NoRels were replaced by an Implicit with a non-adjacent argument, as in Example 3.

(3) *The people in the second circle are the lustful.* Their punishment is to bang against one an another in Hell for all eternity. The modern day examples would be prostitutes or Jerry Sandusky. Next, **The third circle is for the gluttons.** (Implicit/Expansion)

Other potential NoRels were deemed better classified as indirect EntRels (i.e. set/subset, part/whole, or other bridging inferences) (Prince, 1981). However some ambiguity typically remained since EntRels can be extremely indirect in the essays, which also contributes to their lowered coherency. In Example 4, an encompassing entity extending through time can be inferred from "the world today" and "In Dante."

(4) *There are many types of people in the world today, people with different beliefs.*

[4]In the BioDRB, NoRels still occurred in the abstracts and were used to mark duplicate sentences (Prasad et al., 2011).

In Dante, there are different circles for every level of hell. (EntRel)

Considering relation sense, the final row of Table 3 shows the overall percentage of each Level-1 sense for Explicits and Implicits, as computed by totaling all occurrences in every sense combination (e.g., Comp = 315+186+3+9+11 = 524/4262 = 12%). Sense distributions for these types in the PDTB and BioDRB are shown in Table 4.

The essays contain substantially fewer Comparisons than the PDTB (12% versus 23%) but are very similar to the BioDRB, which contains 11%. This suggests Comparisons tend to have less use in argumentative texts, regardless of their level of sophistication. On the other hand, the essays contain substantially more Contingencies (34%) than both the PDTB (22%) and the BioDRB (20%). This may reflect a "sledgehammer" approach to argument construction, and thus a target for learning-based language technology research.

Temporals occur less frequently in the essays than in the PDTB (6% versus 13%) because in the essays most ordering is done in relation to the exposition and so falls under the definition of Expansion, as shown in Example 5. However, the BioDRB contains a much higher proportion of Temporals (17%) that may reflect a more sophisticated use of temporal ordering for argument construction, and another target for learning-based language technology research.

(5) *The fourth level of Hell is the hoarder/ spendthrifts of life. ...* Lastly, **the Wrathful are those who are active while others are passive.** (Explicit/Expansion)

The tendency in the essays to order propositions may also account for the increased proportion of Expansions (48%) as compared to the PDTB (42%). A comparison can't be made here with the BioDRB senses because some map to both Expansion and EntRel (see Footnote 1).

However, the essays' relative proportions of Implicit/Expansion, Implicit/Contingency, and EntRel should be considered fluid, because noise heightened the ambiguity between them. Relation *concurrency* is more common in published texts, i.e. multiple relations holding between two arguments simultaneously (exemplified by "when" and "since," which can convey Contingency and Temporal senses concurrently). Relation *ambiguity* is more common in the essays, however, and particularly between these three relations. EntRels' indirectness is often the cause, exacerbating the ambiguity with Implicit/Expansion even despite the PDTB framework's subdivision of the latter into 10 sub-categories. However, a better explanation of how phrases function as connectives would also help. In Example 6, "In this case" can be inserted but is not listed in the PDTB manual, although other prepositional phrases with abstract objects are, e.g. "as a result," "to this end," etc. If "in this case" is a connective the relation may be an Expansion; else it is probably an EntRel.

(6) *For example an Indian tribe that worships the moon but not God.* **There is no real punishment but the fact that they cannot go to heaven.** (Implicit/Expansion $\vee$ EntRel?)

The ambiguity between Implicit Expansion and Contingency appears partially rooted in the noise of learning. Students are still acquiring the ability to assert causality through voice and language and so their sentences are not always clearly linked. However the ambiguity also results from argument construction. Thus did the BioDRB researchers recognize a need to distinguish two new classes of Contingency: Claims and Justifications, which hold when one situation is the cause for the truth or validity of a *proposition*, from the PDTB's Reasons and Results, which hold when one situation is the cause of another situation. In our data Claims and Justifications often occur with a modal verb, which can disambiguate cases such as Example 7 but not Example 8, suggesting the ambiguity is a function of both noise and domain.

(7) *A hoarder in life would be myself.* Because **I love ice-cream and keep large amounts in my freezer.** (Implicit/Contingency:Justification)

(8) *The descent into the pit of hell would likely be peppered with many more of the faces of todays celebrities.* Because/In other words **Our world today is easily as corrupt as that in which Dante lived.** Sins are timeless, and, in Dantes view, their corresponding punishments are eternal.(Implicit/Expansion $\vee$ Contingency?)

Parser/ Train/ Senses	Test/ Senses	Overall E-to-E pMatch	Overall E-to-E eMatch	Conn ID	Arg ID pMatch	Arg ID eMatch	EXP E-to-E pMatch	EXP E-to-E eMatch	NoEXP E-to-E pMatch	noEXP E-to-E eMatch
Lin14 PDTB L1	**Essays L1**	45%	31%	90%	85%	57%	64%	36%	39%	26%
Lin14 PDTB L2	**Essays L1**	38%	26%	90%	85%	57%	63%	39%	27%	20%
Lin14 PDTB L2	**PDTB L2**	38%	21%	94%	81%	40%	81%	-	25%	-
CoN15 PDTB L2	**PDTB L2**	-	30%	94%	-	49%	-	40%	-	20%
CoN15 PDTB L2	**WikiN L2**	-	24%	92%	-	46%	-	31%	-	19%

Table 5: Comparison of F1 Scores across Discourse Parsers, Training and Test Sets

4 Automatic Discourse Relations

We used the PDTB-trained Lin et al. discourse parser (Lin et al., 2014) to automatically predict our human-annotated relations. As the first end-to-end free text PDTB discourse parser, it is typically the parser to which novel technical advances are compared (e.g., (Xue et al., 2015; Ji and Eisenstein, 2014)). In its sequential pipeline architecture, all functional occurrences of a predefined set of discourse connectives are identified, and then their two arguments are identified and assigned a sense. Subsequently within each paragraph all remaining unannotated adjacent sentence pairs are labeled as Non-Explicit, and their argument spans are identified and assigned a sense. EntRel, AltLex and NoRel relations are also predicted during this step. Since our essays are only annotated with Level-1 senses, we used the Lin et al. parser[5] in two different ways. First, we used the original parser trained on PDTB Level-2 senses to parse essays in terms of Level-2 senses; we then converted the predicted Level-2 senses to their Level-1 abstractions. Second, we retrained the parser by using only PDTB Level-1 senses; this retrained Lin et al. parser directly predicted Level-1 senses.[6]

Table 5 compares both versions of the Lin et al. parser's performance on the essays predicting Level-1 senses, with the original parser's performance on the PDTB test set predicting Level-2 senses. Also compared are variations of the Lin et al. architecture recently evaluated in the CoNLL-2015 Shared Task on Shallow Discourse Parsing (Xue et al., 2015) (**CoNNL15**), trained on and predicting a similar set of Level-2 senses. The fourth row compares the best parsers from this task on the PDTB test set, while the fifth row compares them on the task's own blind test set of WikiNews texts. Note the essays can be viewed as a similar blind test set for the Lin et al. parser, in that the WikiNews texts and essays are unpolished and unpublished; however spelling and grammar errors were removed from the WikiNews texts.

As shown, performance is typically assessed in terms of an F1 score. F1s are computed for overall end-to-end performance (**Overall E-to-E**) as well as performance on the first step of connective identification (**Conn ID**) and the second step (with error propagation from the first step) of argument span identification (**Arg ID**). The F1 score for the final step of sense assignment (with error propagation from the first two steps) corresponds to end-to-end performance. End-to-end performance on Explicits (**EXP E-to-E**) is also distinguished from Non-Explicits (**NoEXP E-to-E**), i.e. Implicit, AltLex and EntRel. Further, within each evaluation (except for the first step of ConnID), performance can be evaluated using exact match (**eMatch**), whereby the parser's arguments must exactly match the human's, or using partial match (**pMatch**), whereby the spans may exactly match or overlap. The CoNLL-2015 Shared Task did not report partial match results even though as Lin et al. (2014) note, most disagreements between exact and partial match do not show significant semantic differences (Miltsakaki et al., 2004) and result from small text portions being included or deleted

[5] `wing.comp.nus.edu.sg/~linzihen/parser`

[6] Thanks to Ilija Ilievski of the National University of Singapore for retraining the Lin et. al parser, and running both the original and retrained versions on our essay corpus.

to enforce the minimal argument constraint, whose presumption of deep semantics poses difficulties for parsers. Because noise made determining minimal arguments problematic (Section 3.1), we report exact and partial match results.

Measuring overall end-to-end performance, Table 5 shows that on the essays the Lin et al. parser yielded F1s of 45% with partial match and 31% with exact match when trained on L1, while its F1s when trained on L2 were lower (38% and 26%). On the PDTB test set its F1s were also lower (38% and 21%). The best CoNLL-2015 parser improved upon the Lin et al. parser for exact match both on the PDTB test set and their own blind test set. Because the annotations being predicted were somewhat different in each case, breaking down performance into component steps helps clarify the import of these results.

On the first step of connective identification, Table 5 shows that performance is uniformly high, which is unsurprising since few explicit connectives are ambiguous (Pitler et al., 2008; Lin et al., 2014; Prasad et al., 2011). On the essays the Lin et al. parser yielded a slightly lower F1 of 90%; this was due to grammatical errors that caused it to miss some connectives, and the fact that it did not recognize all the human-annotated connectives, including prepositional phrases such as "in that case" and "after all." On the second step of argument span identification (with error propagation from connective identification but regardless yet of relation type or sense), Table 5 shows that on the essays the Lin et al. parser yielded partial and exact match F1s of 85% and 57%, outperforming all other parsers and corpora. This was almost certainly because the minimal argument constraint was not strictly enforced in the essay annotation due to noise making argument boundaries ambiguous (Section 3.1); the larger argument enabled more exactly and partially matched spans. Whether relaxing the minimal argument constraint could also increase the usefulness of automatic discourse relation annotation in language technology applications is still an open question.

Finally contrasting end-to-end parser performance on Explicits and Non-Explicits as well as Overall, Table 5 shows the performance improvement on the essays is reduced. In particular, the 8-17% increase over other test sets and parsers for exact match argument identification drops once relation type and sense are predicted for those arguments. Overall the L1 trained essay parser only retains a 1-10% increase, while the L2 trained version's increase is less or nonexistent. Thus even relaxing the minimal argument constraint and predicting only Level-1 senses cannot fully temper the negative impact of noise. Interestingly, the L1 trained essay parser performs better on the Non-Explicits but the L2 trained essay parser performs better on the Explicits; this suggests that the greater training specificity helps to counteract the effect of noise when parsing Explicits.

Table 6 illustrates patterns of errors that occur in the final steps of relation type and sense identification, presenting a confusion matrix of the 4216 relations in the essays whose arguments were at least partially matched. Considering first Explicits, Table 6 shows most disagreements involve parser predictions of Explicit/Temporal (9+28+11) for connectives that can take other senses as well, such as "since" in Example 9 as well as "then" used for textual instead of temporal ordering (Section 3.3). In addition, the parser failed to identify a number of explicit connectives signaling Expansion, labeling them instead as Implicit/Contingency (7) or Implicit/Expansion (22), including sentence-initial, comma-delimited "First" and "Next" as well as sentence-final "too" and "as well." Further investigation is needed to determine why.

(9) *He now has to spend eternity in the second circle of hell* since **he ruined his marriage as a "cheetah" and not a Tiger.** (Human: Explicit/Contingency; Parser: Explicit/Temporal)

Considering Non-Explicits, Table 6 shows no AltLex were predicted by the parser, not surprising since AltLex are so syntactically productive and only the first three stemmed terms of the second argument span were used by the Lin et al. parser to identify them. However, in these essays the human annotator had a highly repetitive cue signaling the most commonly occurring AltLex relation, namely various syntactic permutations of "The example is..." as in Examples 10 and 11. Most of the 99 Implicit/Expansions the parser mislabeled as EntRel contained further permutations of this relation, as shown in Example 11. This suggests that training the parser on essay data could improve its performance on AltLex, EntRel, and Implicit/Expansion.

(10) *Their punishment is to "bang" against*

	Exp: Comp	Exp: Cont	Exp: Expn	Exp: Temp	Imp: Comp	Imp: Cont	Imp: Expn	Imp: Temp	EntR	AltL
Exp: Comp	**191**	1	3	9	0	2	3	0	0	0
Exp: Cont	0	**422**	1	28	0	2	3	0	0	0
Exp: Expn	0	0	**253**	11	0	7	22	0	1	0
Exp: Temp	6	2	1	**140**	0	3	2	0	0	0
Imp: Comp	0	1	0	0	**5**	36	108	0	7	0
Imp: Cont	0	3	0	0	7	**119**	524	2	35	0
Imp: Expn	0	0	1	1	16	208	**1017**	1	99	0
Imp: Temp	0	0	0	0	0	4	16	**0**	0	0
EntR	4	3	5	1	1	121	569	0	**99**	0
AltL	0	1	0	0	1	18	59	0	11	**0**

Table 6: 4216 Partially-Matched Argument Relations in Student Essays (Human: rows, Parser: cols)

one an another in Hell for all eternity. **The modern day examples would be prostitutes or Jerry Sandusky.** (Human: AltLex/Expansion; Parser: EntRel)

(11) *The fourth level of Hell is the hoarder / spendthrifts of life.* As an example, **The person that falls into this layer is Christopher Sisley.** (Human: Implicit/Expansion; Parser: EntRel)

Otherwise Table 6 reflects the relation ambiguity that occurred in the human annotation (Section 3.3). That is, the clusters of counts around the diagonal show the parser also had difficulty distinguishing Implicit/Contingency, Implicit/Expansion and EntRel. As illustration, Example 12 shows one of the 208 cases in which the human annotated Expansion and the parser, Contingency. Example 13 shows one of the 524 cases where the human annotated Contingency and the parser, Expansion. Example 14 shows one of the 569 cases where the human annotated EntRel and the parser, Expansion.

(12) *Pretty much any teenage boy you talk to is gluttonous and never stops eating.* **Every meal is large and overindulgence in food happens every day.** (Human: Implicit/Expansion (In other words); Parser: Implicit/Contingency)

(13) *Paul Fields is one who is in this layer of Hell.* **He scorn the name of band kids who have no idea what they are doing.** (Human: Implicit/Contingency (Because); Parser: Implicit/ Expansion)

(14) *The third circle is for the gluttons.* **They are not only gluttons for food but also gluttons for attention.** (Human: EntRel; Parser: Implicit/Expansion)

Finally, inspection of the 883 remaining disagreed relations (5099-4216) whose arguments weren't both at least partially matched shows as expected that the parser disagreed with 55 Implicits whose left argument was non-adjacent (Section 3.1), since it only labeled Implicits between adjacent sentences. As expected the parser also failed to recognize many relations holding between ungrammatical sentences (Section 3.1), although a manual accounting is still necessary to determine exactly how often this occurred.

5 Conclusions

We investigated manual and automatic PDTB discourse relation annotation in high school student AP English essays. In contrast to prior PDTB applications, the essays are learning-based, in that the writers are learning about argumentative writing through the essay-writing process, and they are also noisy, containing errors of spelling, grammar, and deeper cohesive ties. We discussed methodological complexities of noisy learning-based data, including a heightened ambiguity between EntRel, Expansion, and Contingency that the PDTB framework does not yet resolve. Descriptive statistics showed how relation distributions differ from the PDTB (Prasad et al., 2014) and BioDRB (Prasad et al., 2011) corpora, and also suggested possible targets for future learning-based language technology research. Comparison of automatic discourse parser performances showed that relaxing the minimal argument constraint and predicting only Level-1 senses helped counter the negative impact of noise; the Lin et al. parser, when trained on the PDTB's Level 1 senses, gave an overall F1 score of 31% under strictest evaluation terms, similar to other corpora and parsers (Lin et al., 2014; Xue et al., 2015). Performance was highest on connective and ar-

gument identification, and dropped precipitously during relation type and sense identification. Patterns of errors occurring in those steps indicate training on essay data would improve the parser's ability to distinguish AltLex, Implicit/Expansion, and EntRel, but distinguishing EntRel, Expansion, and Contingency requires first resolving these ambiguities in the manual case. Our results thus support prior work suggesting benefits to tailoring manual annotations to the target data (Zeyrek et al., 2013) and training domain-specific parsers to predict them (Prasad et al., 2011; Ramesh and Yu, 2010).

We are currently exploring the effectiveness of other available discourse parsers. We also plan to annotate and release a new corpus of student essays[7] that we are currently collecting. In addition, we are starting to explore the relationships between student learning and discourse relations, including not only relation use but also the manual and automatic annotations. For example, there may be an interaction such that more coherent, less ambiguous essays also receive higher grades. We will also investigate ways in which annotated discourse relations in learning-based domains can be used to improve existing educational technologies such as language-based tutors and writing assistants (e.g., (Litman and Forbes-Riley, 2014; Zhang et al., 2016)). Level-1 senses have already been shown to be useful for improving sentiment analysis in product reviews (Yang and Cardie, 2014), and we are seeing improvements when using Level-1 senses to enhance our prior work on classifying writing revisions.

Acknowledgments

This work was funded by the Learning Research and Development Center at the University of Pittsburgh.

References

Amal Alsaif and Katja Markert. 2011. Modelling discourse relations for arabic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 736–747, Stroudsburg, PA, USA. Association for Computational Linguistics.

Laurence Danlos, Diego Antolinos-Basso, Chlo Braud, and Charlotte Roze. 2012. Vers le FDTB: French Discourse TreeBank. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, pages 471–478, Grenoble, France.

Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, A. Joshi, B. Webber, Aravind Joshi, and Bonnie Webber. 2002. D-LTAG system: Discourse parsing with a lexicalized tree adjoining grammar. *Journal of Logic, Language and Information*, 12:261–279.

Katherine Forbes-Riley, Bonnie Webber, and Aravind Joshi. 2006. Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, 23(1):55–106.

Sucheta Ghosh, Richard Johansson, Giuseppe Riccardi, and Sara Tonelli. 2012. Improving the recall of a discourse parser by constraint-based post-processing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12); Istanbul, Turkey; May 23-25*, pages 2791–2794.

Yangfeng Ji and Jacob Eisenstein. 2014. One vector is not enough: Entity-augmented distributional semantics for discourse relations. *CoRR*, abs/1411.6699.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Diane Litman and Katherine Forbes-Riley. 2014. Evaluating a spoken dialogue system that detects and adapts to user affective states. In *Proceedings 15th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, Philadelphia, PA, June.

Eleni Miltsakaki, Re Creswell, Katherine Forbes, and Aravind Joshi. 2003. Anaphoric arguments of discourse connectives: Semantic properties of antecedents versus non-antecedents. In *In EACL Workshop on Computational Treatment of Anaphora*.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. In *In Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16.

Gary Patterson and Andrew Kehler. 2003. Predicting the presence of discourse connectives. In *EMNLP*, pages 914–923. ACL.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily Identifiable Discourse Relations. In *Coling*, pages 87–90, Manchester, UK, August. Coling 2008 Organizing Committee.

[7]The existing essays were collected for another project prior to our PDTB research and unfortunately cannot be freely distributed.

Lucie Polkov, Jiri Mrovsk, Anna Nedoluzhko, Pavlna Jnov, Srka Ziknov, and Eva Hajicov. 2013. Introducing the Prague Discourse TreeBank 1.0. In *International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, Japan.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *In Proceedings of LREC*.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind K. Joshi, and Hong Yu. 2011. The Biomedical Discourse Relation Bank. *BMC Bioinformatics*, 12:188.

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

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Syntax and semantics: Vol. 14. Radical Pragmatics*, pages 223–255. Academic Press, New York.

Balaji Polepalli Ramesh and Hong Yu. 2010. Identifying discourse connectives in biomedical text. In *AMIA Annual Symposium Proceedings*, volume 2010, page 657.

T. Sanders and H. Pander Maat. 2006. Cohesion and coherence: Linguistic approaches. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 591–595. Elsevier, Amsterdam.

Himanshu Sharma, Praveen Dakwale, Dipti Misra Sharma, Rashmi Prasad, and Aravind K. Joshi. 2013. Assessment of different workflow strategies for annotating discourse relations: A case study with HDRB. In *Computational Linguistics and Intelligent Text Processing - 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I*, pages 523–532.

Fatemeh Torabi Asr and Vera Demberg. 2013. On the information conveyed by discourse markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 84–93, Sofia, Bulgaria, August. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol T. Rutherford. 2015. The CoNLL-2015 Shared Task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–16, Beijing, China, July.

Bishan Yang and Claire Cardie. 2014. Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 325–335, Baltimore, Maryland, June.

Deniz Zeyrek, Işın Demirşahin, Ayışığı B Sevdik Çallı, and Ruket Çakıcı. 2013. Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue & Discourse*, 4(2):174–184.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–143, Denver, Colorado, June. Association for Computational Linguistics.

Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B. Hashemi. 2016. Argrewrite: A web-based revision assistant for argumentative writings. In *Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL-HLT)*, San Diego, CA, June.

Yuping Zhou and Nianwen Xue. 2015. The Chinese Discourse Treebank: a chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.

A Appendix

A.1 Low Noise Start-of-Essay Excerpt

In Dante's Inferno we have read about the first five circles. Each circle has a different punishment for each sin. In this paper I will fit modern day people into each circle.

A.2 High Noise Start-of-Essay Excerpt

The ones who are born to not flesh nor earth, where blessed with the divine grace and the highway of hell based on Dante's representation of Hell. They are watchers; they have seen Dante's struggles on Earth as well as his teachings through his book. They are all-knowing and represent what Dante tried to explain through his interpretations of Hell. Although he was a bit off they have the true story to be told. To make the levels more relatable they have place modern day people to accompany each level.

A.3 Low Noise Mid-Essay Excerpt

As Dante descends into the second circle, he sees "the sinners who make their reason bond thrall under the yoke of their lust" (98). These were the souls of those who made an act of love, but inappropriately and on impulse. This would be a fine level of Hell for all those who cheat on their boyfriends or girlfriends in high school because let's face it; they aren't really in love.

A.4 High Noise Mid-Essay Excerpt

Michael B calls this home as he was lazy and enjoyed himself to much in his life as a person. Within his home he kept the foods that satisfied his sin, indulging in them whenever he could. The reasoning of this was due to his insatiable appetite, which seemed to never end as he continues to do this sin without much notice and without many hurtles to keep him from the craves. Being housed within the circle he would lay in the mud of waste, living in the waste of the sin that he lives with. While Cerberus acts as his actual sin, him wanting more therefore having three heads. This would give him the experience of the sin that Michael housed within him.

A.5 Essay Examples of Relation Annotations

I don't personally know anyone that is over 2012 years old so **I cannot place any modern people into this layer.** (Explicit/Contingency)

Usually when **I get money** *I plan what I am going to use it for and wait until I have that much or spend it immediately on something I probably don't need.* (Explicit/Temporal)

Filled with hatred for many, yet **never acts upon his grim thoughts.** (Explicit/Comparison)

The man who is stuck in this layer is Hue Heffner. Because **He has devoted his entire life for other people's lustful pleasure and his own.** (Implicit/Contingency)

A prime example of this is a woman by the name of Marie, who abandons man after man in search of a thrill, thrusting her body to anyone willing enough. In other words **She leaves one man for the arms of another, just as Francesca fled to Paolo for satisfaction.** (Implicit/Expansion)

Teachers such as Mr. Braverman are externally wrathful and intentionally cause agony to others like Mrs. Pochiba. In contrast **Other English teachers, such as Mrs. Butler, are very quiet and don't let people know that certain things bother her.** (Implicit/Comparison)

The punishment for these people is to bleed forever with worms sucking up the blood at their feet. **The example would be people who would not choose a side in the civil war.** (AltLex/Expansion)

He does not believe in Christ, but believes in the religion of scientology. **Due to this, he is against the fact that Christ had existed, and had been on Earth.** (AltLex/Contingency)

It may cause you fame and fortune, but what is money if you are greedy? **Although Donald Trump doesn't look at it that way, in God's eyes greed gets you nowhere but the third circle of Hell.** (EntRel/-)

He gives us a better understanding of why certain people are in a certain level of hell. **I will be discussing in the following paragraphs people who deserve to be in each level of hell, in Dante's perspective.** (EntRel/-)

She is young and has not experience a lot of things to be put into a certain level of sin. **The level I'm currently discussing is located in between heaven and hell.** (EntRel/-)