Towards Finer-Grained Tagging of Discourse Connectives

Yannick Versley
Universität Tübingen

Abstract
Many recent experiments in the automatic classification of discourse relations have limited themselves to a small set of coarse categories. While there are eminent reasons to do so – annotation for a small sets of categories can be created more reliably, and possibly also be defined in a more clear – cut way than finer distinctions – it is an interesting question whether the finer-grained distinctions present in some annotated corpora can be reconstructed reliably. The present paper investigates the feasibility of such fine-grained tagging of discourse relations using data from the Penn Discourse Treebank.

1 Introduction
In order to structure a discourse beyond the level of single clauses and the predicate-argument relations contained therein, speakers or writers implicitly or explicitly express relations between events, propositions, or speech acts expressed in different clauses – so called *discourse relations*. Often (but still in a minority of cases) such discourse relations are marked by *discourse connectives*, which signal the presence of a relation between arguments that can be determined either purely syntactically (in the case of coordinating or subordinating conjunctions) or anaphorically (e.g., in the case of discourse adverbials).

Early work on discourse parsing (Soricut and Marcu, 2003) has focused mostly on such overtly marked discourse relations – both because they are easier to detect in general and because the discourse connective itself considerably constrains the kinds of relations that can hold between its arguments. (Some connectives such as *although* always mark one kind of relation, whereas other connectives such as *since* or *and* are more ambiguous).

Later work such as Sporleder and Lascarides (2008); Pitler et al. (2009); Lin et al. (2009) focused on the sense disambiguation of implicit discourse relations, which is more sensitive to semantic information as the lack of an explicit connective yields a significantly higher ambiguity for the realized relation. However, classification accuracy on implicit discourse relations only reaches accuracies of 44.6% (Pitler et al., 2009, for the 4 upper-level categories of the Penn Discourse Treebank (PDTB) plus *EntRel* and *NoRel* for the non-presence of a discourse relation), or 40.2% (Lin et al., 2009, for the 16 mid-level PDTB categories), despite the fact that the (textual/discourse) units to be related are assumed as given. Therefore, methods using explicit cues are currently closer to being useful in actual applications.

Of the existing research on disambiguating the discourse relations signaled by connectives, Haddow (2005) and Miltsakaki et al. (2005) focus on a small number of ambiguous connectives, using a set of relations motivated by said set of connectives. In

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 145-155, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

contrast, Pitler and Nenkova (2009) consider the full range of discourse connectives present in the PDTB, which allows to gain a more comprehensive overall picture. Pitler and Nenkova report results only for the topmost level of the PDTB's relation inventory, which comprises four coarse relation types (*Comparison*, *Expansion*, *Contingency* and *Temporal*). As the PDTB (on the finer granularity levels of their label set) – as well as most other discourse corpora – includes finer distinctions, it may be of interest whether these finer distinctions can also be made automatically. Accurate classfication also at the lower level, using methods that assume only information that can be produced by automatic preprocessing, would clearly also be beneficial from an application perspective.

In this paper, we discuss the problems that are faced by discourse tagging in the finer distinctions of the relation taxonomy, and propose suitable methods for hierarchical classification that allow the prediction of finer classes while making use of the taxonomical information contained in the PDTB's hierarchical label set. We also discuss additional features that help in making these finer-grained distinctions more robustly.

2 Setting

2.1 Data: The Penn Discourse Treebank

The Penn Discourse Treebank 2.0 (PDTB; Prasad et al., 2008) contains, for the text basis covered by the Wall Street Journal portion of the Penn Treebank, annotation of discourse relations marked by a connective (*Explicit*), those that are not marked by a connective (*AltLex* and *Implicit*), as well as annotations that do not signal a discourse relation (*EntRel* and *NoRel*). The present study focuses on the 15 366 *Explicit* relations found in the PDTB (or more specifically, its sections 2-22).

The four coarse relation types in the Penn Discourse Treebank (*Comparison*, *Expansion*, *Contingency* and *Temporal*) are further subdivided into sixteen second-level relations, among which ten occur more than 200 times within sections 2-22: Within the *Comparison* group, this includes the distinction between *Concession* and *Contrast*, within the *Contingency* group, the one between *Cause* and *Condition*, and the *Expansion* group includes, besides *Instantiation* as its predominant member, the *Alternative*, *Instantiation*, and *List* relations. Within the *Temporal* group, a further distinction is drawn between *Asynchronous* and *Synchrony* relations. (Among the infrequent second-level relations are the 'pragmatic' variants of concession, contrast, cause, and condition, all occurring 50 times or less, as well as *Expansion.Restatement*, which occurs 128 times).

The third level, finally, distinguishes multiple variants of *Contrast* and *Concession* based on the relation between the objects or propositions that are related, *Cause* contains the division between *Reason* and *Result* (corresponding to the causal ordering of the assigned arguments, which normally only varies with syntactic properties of

the connective), as well as various distinctions among *Condition* based on factuality; within the *Asynchronous* temporal relations, the third level distinguishes *Precedence* and *Succession*.

2.2 Features

Discourse classification is carried out as a supervised machine learning task, using features that summarize the linguistic properties of the discourse connective's context. Two of the used features are reimplementations of ones used by Pitler and Nenkova (2009): One is the string of the connective itself. In order to reduce the annotated spans in the *connector* slot of the annotation, which can include additional text to the connective itself (such as "*two minutes*" in "*two minutes before the train departed*"), occurrences ending with one of *after*, *before*, *when*, *until*, *since*, or *if* were shortened to that word whenever the span was longer. The case of all connectives was normalized to lower case.

The second group of features comprises Pitler and Nenkova's syntactic features: These include the labels of self, parent, left sibling and right sibling nodes (counting from the lowest node that covers all of the words annotated as connective span and that is not the only child of its parent), as well as additional features signaling the presence of a VP node or of a trace as a child of the right sibling.

In line with the observations by Pitler and Nenkova, we found that the ambiguity between *Temporal* relations and *Contingency* relations (specifically, *Condition*) was a major source of misclassifications. The main difference between *Temporal* and *Contingeny* relations in the explicit cases lie in the facticity of the connected events. Both Miltsakaki et al. (2005) and Haddow (2005) use additional features that pertain to tense and mood of the connected arguments, but presuppose the arguments as given.

To be able to use these features with automatic preprocessing and tell whether they are informative with respect to the distinction between *Temporal* and *Contingency* relations (as well as the accuracy of relations on the finer levels of the taxonomy), we automatically derive the argument nodes from the syntactic annotation of the treebank. While the PDTB annotation contains argument spans, methods for their automatic identification are not perfect – Elwell and Baldridge (2008) report accuracy scores of 82.0% and 93.7%, which means that using perfect information in the identification of discourse relations may create a distorted picture.

As a simple, high-precision mechanism to identify arguments, we implemented heuristics to derive the argument nodes using syntactic heuristics for different groups of connectives, in particular subordinating coordinators (*[S . . . [SBAR [IN after] she slept]]*), clausal PPs (*[S . . . [PP [IN after] [S sleeping]] . . .]*), sentence coordination (*[S [S he sleeps] [CC and] [S he snores]]*), w-adverbials (*[S . . . [SBAR [WHADVP when] he sleeps]]*), as well as fronted (preposition- or adverb-headed) adverbials, which have

one of their arguments (the ARG2 in PDTB parlance) in the current sentence whereas the other is linked anaphorically.

Based on the identified arguments, we extract the following indicators:

- the part-of-speech of the first non-modal verb in the sentence (descending from the argument clause node into further VP and S nodes to cover both nesting of VPs and coordinated sentences)
- the presence (and word form) of modals and negation in the clause
- a tuple of *(have-form, be-form, head-POS, modal present)* as proposed by Miltsakaki et al. (2005).

(In the result tables, the part-of-speech/presence of modals pair of features will be called *pos*, whereas the tuple describing auxiliaries, the POS of the lexical head, and the presence of modals will be simply called *verb*).

Verb tense and modals are relatively shallow correlates of more interesting properties such as facticity or veridicality (i.e., whether the speaker asserts the propositional content of that clause to be true), but they are easy to extract in a robust manner and useful as a first approximation to a more comprehensive approach such as those of Palmer et al. (2007) to classifying situation entities.

2.3 Hierarchical Classification

Considering that the Penn Discourse Treebank has a hierarchical label set, relevant generalizations may be found at multiple levels of the relation hierarchy. In the area of word sense disambiguation, Ciaramita et al. (2003) have shown that a classifier that uses a two-level hierarchy to generalize the word senses performs better than a state-of-the-art "flat" multiclass classifier.

For our version of the hierarchical classification, we start from a maximum entropy classifier (Berger et al., 1996), in contrast to Ciaramita et al., who use a Perceptron classifier.[1] In the standard formulation, maximum entropy learning minimizes the loss

$$\mathrm{Loss}(w) = \prod_{x,y} \log \frac{\mu(x,y)}{\sum_{y' \in Y} \mu(x,y')}$$

where the measure $\mu(x,y)$ is defined as

$$\mu(x,y) = \exp(\langle w, \phi(x,y) \rangle)$$

for a feature function ϕ that pairs all features extracted from x with the label for y.

In the hierarchical case, ϕ pairs the features extracted from x not only with the actual class label y, but also nodes higher up in the taxonomy - yielding, for example, not only

[1] A wide variety of learning algorithms can be used to learn linear multiclass classifiers such as those used by Ciaramita et al. and in this work, of which the standard techniques for maximum entropy estimation – optimizing a log-likelihood-based loss using quasi-Newton numerical optimization – are by far the most commonly used.

a weight for "x has an SBAR parent and y is *Contingency.Cause.Result*", but also for the more general "x has an SBAR parent and y is a descendent of *Contingency.Cause*". To improve the separability of the problem at hand, we consider combinations of up to two of the original features from x.

As the PDTB contains underspecified relations (e.g., just *Contingency*) in cases where annotators could not reach an agreement about the finer relation, such labels would occur as possible tags for relation instances, including those that are tagged with a finer label. To avoid the confusion that would arise from using the underspecified relations either as positive or negative example, we completely removed the less-specific relation from the learning instance if it was labeled with a more-specific relation.

To make use of the presence of multiple relation labels in the annotation of the Penn Discourse Treebank (for example, a given instance of a connective may receive *Temporal.Synchrony* as the primary classification and *Comparison.Contrast.Juxtaposition* as a secondary classification) we chose to optimize the (sum) probability that the model assigns to *all* of the correct labels:

$$\text{Loss}(w) = \prod_{x, Y_{good}} \frac{\sum_{y \in Y_{good}} \mu(x, y')}{\sum_{y' \in Y} \mu(x, y')}$$

Besides the flat multiclass classifier and the hierarchical classifier, we also implemented a method for *greedy* classification, where the top-level relation is determined and subsequent relations are determined by a specialized classifier that, for a given relation prefix, guesses the next element. For example, the topmost classifier would classify the relation as *Temporal*, then the second-level classifier for *Temporal* would determine that the relation is *Temporal.Asynchronous*, and the third-level classifier for *Temporal.Asynchronous* would choose *Temporal.Asynchronous.Precedence* as the finest-level relation.

3 Results

For the quantitative evaluation, we follow Pitler and Nenkova in treating the system classification as correct whenever it matches the label, or one of multiple assigned labels, from the manual annotation. To account for the underspecified relations in the PDTB, we also count the system response as correct when it is more specific than the gold-standard label (or one of the gold-standard labels) – for example, when the corpus annotation contains an underspecified *Comparison* annotation, but the system predicts *Comparison.Concession* or even *Comparison.Concession.Contraexpectation*, our evaluation would count this as correct.

Tables 1, 2, and 3 show the results for using different classification methods. Except for the 'greedy' classifier on the finer relations using the approximate tense/mood features, we see only very small differences on the order of 0.1-0.2%, with the greedy classifier performing slightly better on the finer relation levels.

	connective			conn+syntax			conn+verb(arg1)		
evaluated	1	2	3	1	2	3	1	2	3
d=1	0.946	0.946	0.946	0.954**	0.954**	0.954**	0.953**	0.952**	0.952**
d=2		0.840	0.839		0.847*	0.847**		0.845	0.845
d=3			0.790			0.796*			0.798**

Table 1: Flat classification

	connective			conn+syntax			conn+verb(arg1)		
evaluated	1	2	3	1	2	3	1	2	3
d=1	0.946	0.946	0.945	0.954**	0.953**	0.954**	0.953**	0.952**	0.952**
d=2		0.840	0.839		0.847**	0.847*		0.845	0.845*
d=3			0.790			0.796*			0.798**

Table 2: Hierarchical classification

	connective			conn+syntax			conn+verb(arg1)		
evaluated	1	2	3	1	2	3	1	2	3
d=1	0.946	0.946	0.946	0.955**	0.954**	0.955**	0.953**	0.953**	0.953**
d=2		0.840	0.840		0.847*	0.847*		0.845	0.845
d=3			0.792			0.798*			0.800*

Table 3: Greedy classification

Differences to connective-only version: significant at $p < 0.01$ (*) / significant at $p < 0.001$ (**)

Relation	N	Prec	Recl	F
Comparison	**4566**	*0.960*	*0.968*	*0.964*
Comparison.Contrast	3102	0.771	0.898	0.829
Comparison.Concession	1080	0.549	0.309	0.396
Contingency	**2634**	*0.970*	*0.873*	*0.919*
Contingency.Cause	1456	0.982	0.868	0.921
Contingency.Condition	1123	0.919	0.883	0.901
Expansion	**5206**	*0.979*	*0.960*	*0.969*
Expansion.Conjunction	4293	0.920	0.955	0.920
Expansion.Alternative	300	0.926	0.914	0.920
Expansion.Instantiation	245	0.992	0.963	0.977
Expansion.List	205	0.000	0.000	0.000
Temporal	**2961**	*0.882*	*0.966*	*0.923*
Temporal.Asynchronous	1712	0.938	0.869	0.902
Temporal.Synchrony	1244	0.691	0.937	0.795

Table 4: Results for the most frequent second-level relations (connective+syntax)

As can be seen in Table 4, the only problem at the coarsest level of relations is a misclassification of *Contingency* relations as *Temporal*, often in cases such as (1)[2] where the facticity of the sentence cannot be judged without context:

(1) But **when** market interest rates move up rapidly, increases in bank CD yields sometimes lag.

Among the second-level relations, performance on most relations is generally good, with most frequent relations having an F-measure of more than 0.9, but several relations are frequently misidentified: The distinction between *Concession* and *Contrast* – obviously a relatively central one, which however depends on the semantic content of the connective arguments – cannot always be made reliably, and *Concession* as the less frequent relation shows low precision and recall. Within the *Expansion* relations, the lower-frequency relations (of which only *List* is shown) are never predicted because the features used are not strong enough to overcome the preference for the more frequent relations. Within the *Temporal* relations, we see that the effect of misclassifications such as in example (1) is more predominant on the *Temporal.Synchrony* relations.

Pitler and Nenkova's features use the Penn Treebank in its original form, including, on one hand, traces, and, on the other hand, function labels which indicate temporal (−TMP), purpose (−PRP) or other adverbial modification (−ADV).[3] Given an automatic parse, this information would have to be reconstructed, since parsing models are always trained on a version of the treebank that has traces and such semantic function labels removed.[4] Furthermore, the reconstruction of traces and function labels is somewhat error-prone (Gabbard et al., 2006 report an F-measure of about 85% for semantic function tags, and 75% for traces) which means that using this information in sense prediction is prone to overestimating the actual performance in a complete system.

To quantify the influence of this additional gold-standard information, we compute a variant of the syntax features where the trace feature is not used and function labels are stripped from the nodes. As can be seen in Table 5, this version of the syntactic features gives results that are very close to the results that one gets with only the string of the connective.

While the inclusion of tense information cannot improve over the information contained in the semantic function tags (see the *conn+syntax*^{A} and *conn+syntax*^{A}*+tense* rows in Table 5), the incorporation of tense/mood information on the heuristically determined ARG1 (if present in the same sentence) yields useful results by itself.

In contrast, including a single feature that summarizes the syntactic environment (subordinating coordinator, clausal PP, sentence coordination, etc.) and tense features *for the modifiee (Arg1) only* yields results that are close to those with semantic function tags.

[2]The example is annotated as *Contingency.Condition.Hypothetical*, but predicted as *Temporal.Synchrony*

[3]It is not clear from Pitler and Nenkova's paper whether they used a version with function labels or without, since they do not mention it; as they use traces, the most plausible interpretation is that they used a version where function labels are intact.

[4]The −TMP function label on noun phrases is usually kept, since it reflects a syntactic distinction – adverbial versus argument role of the NP – rather than a semantic one and is useful for the parser itself.

	d=1	d=2	d=3
hierarchical			
connective only	0.946	0.839	0.790
conn+syntax^A	0.954	0.847	0.796
conn+syntax^B	0.945	0.840	0.788
w/traces	0.948	0.843	0.792
w/function tags	0.954	0.847	0.796
conn+verb(arg1)	0.952	0.845	0.798
conn+syn^B+pos(arg1)	0.949	0.843	0.794
conn+pos(both)	0.949	0.843	0.794
conn+syn^B+pos(both)	0.947	0.839	0.788
greedy			
connective only	0.946	0.840	0.792
conn+syntax^A	0.955	0.847	0.798
conn+verb(arg1)	0.953	0.845	0.800

syntax^A: with traces and function tags syntax^B: without traces or function tags

Table 5: Different versions of syntactic and tense/mood features

Both the results for syntax including semantic function tags and those for the inclusion of Arg1-related verb features yield improvements over the connective-only version that are statistically significant according to a paired t-test. (All are significant at the $p < 0.05$ level; the improvements on the first level yield p-values around 10^{-5}).

Tables 6 and 7 summarize the behavior of relation prediction over several connectives. Besides the fact that the more difficult task of distinguishing relations according to the larger set also leads to more connectives showing ambiguities, we see that, firstly, the distinction between topicalized and non-topicalized adjunct clauses (summarized as a feature indicating whether the connective is at the start of a sentence, in the column named *conn+first*) has relatively limited benefits. Secondly, the actual syntactic features (*conn+syn^B*, without semantic function labels) and the tense/mood-based features (*conn+mood*) are useful in the case of different connectives – "*since*", for example, does not benefit much from syntactic features but shows a strong improvement when tense and mood information is added.

4 Conclusion

In this paper, we presented first results on the classification of discourse relations using a novel approach that makes use of the hierarchical structure of the label set of the Penn Discourse treebank, and provided an error analysis that extends to the lower levels of

connective	frequency	conn	conn+first	conn+syn^B	conn+syn^A	conn+verb(arg1)
since	154	0.571	0.571	0.675	**0.935**	0.909
finally	30	0.633	0.933	0.867	0.867	**0.933**
in turn	27	0.704	0.704	0.704	0.704	0.704
even as	11	**0.727**	0.636	0.364	0.455	0.636
while	652	0.729	0.727	0.729	**0.839**	0.805
as	588	0.786	0.786	0.781	**0.810**	0.781
as long as	20	**0.800**	0.786	0.750	0.700	0.750

Connectives that occur at least 10 times and have at most 80% accuracy

Table 6: Ambiguous connectives at the coarsest level

connective	frequency	conn	c+first	c+syn^B	c+syn^A	c+arg1	c+syn^B+arg1
rather	14	0.286	**0.643**	0.429	0.357	**0.643**	0.500
as soon as	17	0.294	0.294	0.294	0.176	**0.412**	0.176
nevertheless	30	0.300	**0.533**	0.300	0.333	0.300	0.333
in fact	70	0.300	0.386	0.286	0.300	0.343	**0.429**
finally	30	0.367	**0.667**	0.633	0.533	**0.667**	**0.667**
although	277	0.498	0.588	0.520	0.549	0.592	**0.606**
still	156	0.500	0.429	0.462	**0.506**	0.417	0.449
since	154	0.571	0.571	0.669	**0.929**	0.903	0.896
though	187	0.588	**0.652**	0.540	0.551	**0.652**	**0.652**
while	652	0.598	0.598	0.604	**0.718**	0.667	0.672
indeed	86	**0.605**	0.593	0.593	0.593	0.570	0.558
when	837	**0.611**	0.608	0.609	0.609	0.596	0.588
in particular	13	**0.615**	0.538	0.538	0.538	0.538	0.538
yet	88	**0.648**	**0.648**	0.523	0.432	0.523	0.545
overall	10	**0.700**	0.600	0.400	0.300	0.600	0.400
in turn	27	0.704	0.704	0.704	0.704	0.704	0.704
even as	11	**0.727**	0.636	0.182	0.273	0.636	0.273
as	588	0.745	0.745	0.736	**0.767**	0.743	0.745
in the meantime	12	0.750	0.833	**1.000**	**1.000**	0.833	**1.000**
but	2767	**0.790**	**0.790**	0.789	0.788	0.789	0.785
nor	24	**0.792**	**0.792**	0.667	0.667	0.750	0.667
meanwhile	160	**0.800**	**0.800**	**0.800**	0.775	0.794	0.794
as long as	20	**0.800**	0.750	0.750	0.700	0.750	0.750
ultimately	17	**0.882**	0.765	0.706	0.647	0.765	0.706
now that	20	0.900	0.900	0.550	**0.901**	0.900	0.750

Connectives that occur at least 10 times and have at most 80% accuracy

Table 7: Ambiguous connectives at the medium level

the label hierarchy.

Considering the purpose of applying discourse tagging to raw text, it would be desirable to achieve the tagging of connectives at the granularity of second-level, rather than top-level categories in the Penn Discourse Treebank's inventory, since many important distinctions (*Contrast* versus *Concession*, or *Cause* versus *Condition*) are only made at the second level of the taxonomy. For many of these finer distinctions, neither Pitler and Nenkova's syntactic features nor the tense/mood based that we presented here are sufficient to reach high (>90%) accuracies, despite Pitler and Nenkova's encouraging results on the coarser top-level relation categories.

One plausible reason for this is that the shallow information used by current approaches is not sufficient to reproduce the more semantic distinctions on the finer levels of the taxonomy. Another plausible reason, which has also been pointed out by Pitler and Nenkova concerning the coarser-level distinctions and which we cannot exclude at this point, would be that system accuracy is bounded by annotator agreement: Some distinctions among those in the Penn Discourse Treebank are hard to make reliably even for humans, and similarly our results come close to the levels of annotator agreement reported by Prasad et al. (2008) for the PDTB – 84.5% for the second level, against 84% agreement, and 79.5% for the third level, compared to an agreement figure of 80%.

Our evaluation on the finer levels of the relation taxonomy, however, is slightly more lenient than the annotation in the Penn Discourse Treebank: as we allow any subtype for the underspecified relations where annotators disagreed on the finer relations, these disagreement cases are mostly counted as correct, whereas counting a more specific label as wrong (which would mean that the majority of such disagreement cases would be counted as a disagreement between system and gold annotation, since the system only very rarely assigns an underspecified label) would yield markedly lower results of about 68% for the third-level relations, which would allow for hope of further improvement through more semantic features.

References

Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

Massimiliano Ciaramita, Thomas Hofmann, and Mark Johnson. Hierarchical semantic classification: Word sense disambiguation with world knowledge. In *18th International Joint Conference on Artificial Intelligence (IJCAI 2003)*, 2003.

Robert Elwell and Jason Baldridge. Discourse connective argument identification with connective specific rankers. In *Proceedings of ICSC-2008*, 2008.

Ryan Gabbard, Mitchell Marcus, and Seth Kulick. Fully parsing the Penn Treebank. In *HLT/NAACL 2006*, 2006.

Barry Haddow. Acquiring a disambiguation model for discourse connectives. Master's thesis, School of Informatics, University of Edinburgh, 2005.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *EMNLP 2009*, 2009.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Experiments on sense annotations and sense disambiguation of discourse connectives. In *TLT 2005*, 2005.

Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. A sequencing model for situation entity classification. In *ACL 2007*, 2007.

Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In *ACL 2009 short papers*, 2009.

Emily Pitler, Annie Lous, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *ACL-IJCNLP 2009*, 2009.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL-2003)*, 2003.

Caroline Sporleder and Alex Lascarides. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369–416, 2008.