# Easily Identifiable Discourse Relations

**Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, Aravind Joshi**
University of Pennsylvania
3330 Walnut Street
Philadelphia, PA 19104

## Abstract

We present a corpus study of local discourse relations based on the Penn Discourse Tree Bank, a large manually annotated corpus of explicitly or implicitly realized relations. We show that while there is a large degree of ambiguity in temporal explicit discourse connectives, overall connectives are mostly unambiguous and allow high-accuracy prediction of discourse relation type. We achieve 93.09% accuracy in classifying the explicit relations and 74.74% accuracy overall. In addition, we show that some pairs of relations occur together in text more often than expected by chance. This finding suggests that global sequence classification of the relations in text can lead to better results, especially for implicit relations.

## 1 Introduction

Discourse relations between textual units are considered key for the ability to properly interpret or produce discourse. Various theories of discourse have been developed (Moore and Wiemer-Hastings, 2003) and different relation taxonomies have been proposed (Hobbs, 1979; McKeown, 1985; Mann and Thompson, 1988; Knott and Sanders, 1998). Among the most cognitively salient relations are causal (contingency), contrast (comparison), and temporal.

Very often, the discourse relations are *explicit*, signaled directly by the use of appropriate discourse connectives:



**(E1)** He is very tired **because** he played tennis all morning.

**(E2)** He is not very strong, **but** he can run amazingly fast.

**(E3)** We had some tea in the afternoon and **later** went to a restaurant for a big dinner.

Discourse relations can also be *implicit*, inferred by the context of the utterance and general world knowledge.

**(I1)** I took my umbrella this morning. [**because**] The forecast was rain in the afternoon.

**(I2)** She is never late for meetings. [**but**] He always arrives 10 minutes late.

**(I3)** She woke up early. [**afterward**] She had breakfast and went for a walk in the park.

An additional complication for automatic classification of discourse relations is that even in the presence of an explicit discourse connective, the connective might be ambiguous between several senses. For example, *since* can be used to signal either a temporal or a contingency relation.

They have not spoken to each other **since** they argued last fall. (Temporal)

I assumed you were not coming **since** you never replied to the invitation. (Causal)

Several questions directly related to efforts in automatic recognition of discourse relations arise:

*In a general text, what is the proportion of explicit versus implicit relations?* Since implicit relations are presumably harder to recognize automatically, the larger their proportion, the more difficult the overall prediction of discourse relation will be.

*How ambiguous are discourse connectives?* The degree of ambiguity would give an upper bound on the accuracy with which explicit relations can be identified. The more ambiguous discourse connectives are, the more difficult it would be to automatically decide which discourse relation is expressed in a given sentence, even in the presence of a connective.



*Coling 2008: Companion volume – Posters and Demonstrations,* pages 87–90
Manchester, August 2008

*In a text, are adjacent discourse relations independent of each other or are certain sequences of relations more likely?* In the latter case, a "discourse grammar" of text can be used and easy to identify relations such as unambiguous explicit relations can help determine the class of implicit relations that immediately follow or precede them.

In this study, we address the above questions using the largest existing corpus manually annotated with discourse relations—the Penn Discourse Tree Bank (Prasad et al., 2008).

## 2 The Penn Discourse Tree Bank

The Penn Discourse Treebank (PDTB) is a new resource (Prasad et al., 2008) of annotated discourse relations. The annotation covers the same 1 million word Wall Street Journal (WSJ) corpus used for the Penn Treebank (Marcus et al., 1994).

he PDTB is the first corpus to systematically identify and distinguish explicit and implicit discourse relations. By definition, an *explicit relation* is triggered by the presence of a discourse connective which occurs overtly in the text. The discourse connective can essentially be viewed as a discourse-level predicate which takes two clausal arguments. For example, sentence E1 above could be represented as BECAUSE("He is very tired", "he played tennis all morning"). The corpus recognizes 100 such explicit connectives and contains annotations for 19,458 explicit relations [1].

The PDTB also contains provisions for the annotation of *implicit discourse relations* between adjacent sentences which are inferred by the reader but are not overtly marked by a discourse connective. In this case, the annotator was asked to provide a connective that best captured the inferred relation. There are a total of 16,584 implicit relations annotated in the corpus. [2]

In addition to discourse relations and their arguments, the PDTB also provides the *senses* of each relation(Miltsakaki et al., 2008). The tagset of senses is organized hierarchically into three levels - class, type, and subtype. The top class level contains the four major semantic classes: Expansion, Comparison, Contingency and Temporal.

[1]The PDTB allows annotators to tag a relation with multiple senses. In this work we count both of the annotated senses. So even though there are only 18,459 explicit relations, there are 19,458 explicit senses.

[2]Again, because of multiple senses per relation, the 16,584 senses are part of 16,224 relations.

| Class | Explicit (%) | Implicit (%) | Total |
|---|---|---|---|
| Comparison | 5590 (*69.05%*) | 2505 (30.95%) | 8095 |
| Contingency | 3741 (46.75%) | 4261 (53.25%) | 8002 |
| Temporal | 3696 (*79.55%*) | 950 (20.45%) | 4646 |
| Expansion | 6431 (42.04%) | 8868 (57.96%) | 15299 |

Table 1: Discourse relation distribution in semantic and explicit/implicit classes in the PDTB

## 3 Distribution and ambiguity of connectives

Table 1 shows the distribution of discourse relations between the four main relation classes and their type of realization (implicit or explicit). Interestingly, temporal and comparison relations are predominantly explicit. About 80% and 70%, respectively, of their occurrences are marked by a discourse connective. The contingency relations are almost evenly distributed between explicit and implicit. The expansion relations, the overall largest class of discourse relations, is in most cases implicit and not marked by a discourse connective.

Given the figures in Table 1, we would expect that overall temporal and comparison relations will be more easily identified since they are overtly marked. Of course this would only be the case if discourse markers are mostly unambiguous.

Here we show all connectives that appear more than 50 times in the PDTB, their predominant sense (comparison, contingency, temporal or expansion), as well as the percentage of occurrences of the connective in its predominant sense. For example the connective *but* has *comparison* as its predominant sense and 97.19% of the 3,308 occurrences of this connective were comparisons.

**Comparison** *but* (3308; 97.19%), *while* (781; 66.07%), *however* (485; 99.59%), *although* (328; 99.70%), *though* (320; 100.00%), *still* (190; 98.42%), *yet* (101; 97.03%)

**Expansion** *and* (3000; 96.83%), *also* (1746; 99.94%), *for example* (196; 100.00%), *in addition* (165; 100.00%), *instead* (112; 97.32%), *indeed* (104; 95.19%), *moreover* (101; 100.00%), *for instance* (98, 100.00%), *or* (98; 96.94%), *unless* (95; 98.95%), *in fact* (82; 92.68%) *separately* (74; 100.00%)

**Contingency** *if* (1223; 95.99%), *because* (858, 100.00%), *so* (263; 100.00%), *since* (184; 52.17%), *thus* (112; 100.00%), *as a result* (78; 100.00%)

**Temporal** *when* (989; 80.18%), *as* (743; 70.26%), *after* (577; 99.65%), *then* (340; 93.24%), *before* (326; 100.00%), *meanwhile* (193; 48.70%), *until* (162; 87.04%), *later* (91; 98.90%), *once* (84; 95.24%)

The connectives that signal comparison and contingency are mostly unambiguous. Obvious exceptions are two of the connectives that are often used to signal temporal relations: *while* and *since*.

The predominant senses of these connectives are comparison (66.07%) and contingency (52.17%) respectively. Disambiguating these problematic connectives has already been addressed in previous work (Miltsakaki et al., 2005), but even the predominantly temporal connectives are rather ambiguous. For example less than 95% of the occurrances of *meanwhile, as, when, until,* and *then* are temporal relaions.

While some connectives such as "since" are ambiguous, most are not. The discourse connectives in the corpus appear in their predominant sense 93.43% (for comparsion), 94.72% (for contingency), 84.10% (for temporal), and 97.63% (for expansion) of the time. Temporal connectives are most ambiguous and connectives signaling expansion are least ambiguous.

## 4 Automatic classification

The analyses in the previous section show two very positive trends: many of the discourse relations are explicitly marked by the use of a discourse connective, especially comparison and temporal relations, and discourse connectives are overall mostly unambiguous. These facts would suggest that even based only on the connective, classification of discourse relations could be done well for all data (including both implicit and explicit examples) and particularly well for explicit examples alone. Indeed, Table 2 shows the performance of a decision tree classifier for discourse relations, on all data and on the explicit subset in the second and third column respectively.

We use the natural distribution of relation classes found in the Wall Street Journal texts, without downsampling to get balanced distribution of classes. There are four task settings, distinguishing each type of relation from all others. For example, comparison relations can be distinguished from all other relations in the corpus with overall accuracy of 91.28%, based only on the discourse connective (first line in Table 2). The recall for recognizing comparison relations is 0.66, directly reflecting the fact that 31% of all comparison relations are implicit (Table 1) and the connective feature did not help at all in those cases. Over explicit data only, the classification accuracy for comparison relation versus any other relation is 97.23%, and precision and recall is 0.95 and above.

As expected, the overall accuracy of identifying contingency and expansion relations is lower, 84.44% and 77.51% on all data respectively, reflecting the fact that these relations are often implicit. But by themselves these accuracy numbers are actually reasonable, setting a rather high baseline for any more sophisticated method of classifying discourse relations. On explicit data only, the two-way classification accuracy for the four main types of relations is 94% and higher.

| Task | All relations | Explicit relations only |
|---|---|---|
| Comparison | 91.28% (76.54%) | 97.23% (69.72%) |
| Contingency | 84.44% (76.81%) | 93.99% (79.73%) |
| Temporal | 94.79% (86.54%) | 95.4% (79.98%) |
| Expansion | 77.51% (55.67%) | 97.61% (65.16%) |

Table 2: Decision tree classification accuracy using only the presence of connectives as binary features. The majority class is given in brackets.

| Class | Precision | Recall |
|---|---|---|
| Temporal | 0.841 [0.841] | 0.729 [0.903] |
| Expansion | 0.658 [0.973] | 0.982 [0.957] |
| Contingency | 0.948 [0.947] | 0.369 [0.844] |
| Comparison | 0.935 [0.935] | 0.671 [0.971] |

Table 3: Four-way classification. The first number is for all data, thesecond for explicit relations only.

In four-way classification, disambiguating between the four main semantic types of discourse relations leads to 74.74% classification accuracy. The accuracy for four-way classification of explicit relations is 93.09%. The precision and recall for each class is shown in Table 4. The worst performance on the explicit portion of the data is the precision for temporal relations and the recall for contingency relations, both of which are 0.84.

## 5 N-gram discourse relation models

We have shown above that some relations, such as comparison, can be easily identified because they are often explicit and are expressed by an unambiguous connective. However, one must build a more subtle automatic classifier to find the implicit relations. We now look at the frequencies in which various relations are adjacent in the PDTB. Results from previous studies of discourse relations suggest that the *context* of a relation can be helpful in disambiguating the relation (Wellner et al., 2006). Here we identify specific dependencies that exist between sequences of relations.

We computed $\chi^2$ statistics to test the independence of each pair of relations. The question is: do relations A and B occur adjacent to each other more than they would simply due to chance? The

| First Relation | Second Relation | $\chi^2$ | $p$-value |
|---|---|---|---|
| E. Comparison | I. Contingency | 20.1 | .000007 |
| E. Comparison | E. Comparison | 17.4 | .000030 |
| E. Comparison | I. Expansion | 9.91 | .00161 |
| I. Temporal | E. Temporal | 9.42 | .00214 |
| I. Contingency | E. Contingency | 9.29 | .00230 |
| I. Expansion | E. Expansion | 6.34 | .0118 |
| E. Expansion | I. Expansion | 5.50 | .0191 |
| I. Contingency | E. Comparison | 4.95 | .0260 |

Table 4: $\chi^2$ results for pairs of relations

pairs of implicit and explicit relations which have significant associations with each other ($pval < 0.05$) are shown in Table 4. For example, explicit comparison and implicit contingency co-occur much more often than would be expected if they were independent. As explicit comparisons are generally fairly easy to identify, knowing that they tend to co-occur may be helpful when searching for implicit contingency relations in a text.

## 6 Conclusion

We have tried to summarize the difficulty of finding discourse relations using the Penn Discourse Treebank. We noted that explicit and implicit relations are approximately evenly distributed overall, making the task easier than many researchers have feared. We have found that some relations, such as temporal and comparison, are more likely to be explicit than implicit, making them relatively easier to find, while contingency and expansion are more often implicit. Among the discourse connectives, the majority are not very ambiguous between the different types of relations, with some notable exceptions such as *since* and *meanwhile*.

We have carried out a novel quantitative study of the patterns of dependencies between discourse relations. We found that while there does not appear to be a clear template for the sequence of relations, there are individual relation pairs that tend to co-occur. Specifically, we found that even though contingency relations are likely to be implicit and thus difficult to find, they are likely to be found near an explicit comparison. We plan to exploit these findings in future work, addressing discourse relation labeling in text as a sequence labeling problem and using the explicit cue words of surrounding relations as features for finding the "hidden" implicit relations.

## 7 Acknowledgments

This work was partially supported by an Integrative Graduate Education and Research Traineeship grant from National Science Foundation (NSF-IGERT 0504487) and by NSF Grant IIS -07-05671. We would like to thank Nikhil Dinesh for help with the PDTB, and Rashmi Prasad, Bonnie Webber and Eleni Miltsakaki for insightful discussions.

## References

Hobbs, J. 1979. Coherence and coreference. *Cognitive Science*, 3:67–90.

Knott, A. and T. Sanders. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2):135–175.

Mann, W. and S. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Marcus, M.P., B. Santorini, and M.A. Marcinkiewicz. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

McKeown, Kathleen R. 1985. *Text generation: using discourse strategies and focus constraints to generate natural language text.* Cambridge University Press, New York, NY, USA.

Miltsakaki, E., N. Dinesh, R. Prasad, A. Joshi, and B. Webber. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005).*

Miltsakaki, Eleni, Livio Robaldo, Alan Lee, and Aravind Joshi. 2008. Sense annotation in the penn discourse treebank. *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, 4919:275–286.

Moore, J. and P. Wiemer-Hastings. 2003. Discourse in computational linguistics and artificial intelligence.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC).*

Wellner, B., J. Pustejovsky, C. Havasi, R. Sauri, and A. Rumshisky. 2006. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of the 7th SIGDIAL Workshop on Discourse and Dialogue*.