Realization of Discourse Relations by Other Means: Alternative Lexicalizations

Rashmi Prasad and **Aravind Joshi**
University of Pennsylvania
rjprasad,joshi@seas.upenn.edu

Bonnie Webber
University of Edinburgh
bonnie@inf.ed.ac.uk

Abstract

Studies of discourse relations have not, in the past, attempted to characterize what serves as evidence for them, beyond lists of frozen expressions, or markers, drawn from a few well-defined syntactic classes. In this paper, we describe how the lexicalized discourse relation annotations of the Penn Discourse Treebank (PDTB) led to the discovery of a wide range of additional expressions, annotated as *AltLex* (*alternative lexicalizations*) in the PDTB 2.0. Further analysis of AltLex annotation suggests that the set of markers is open-ended, and drawn from a wider variety of syntactic types than currently assumed. As a first attempt towards automatically identifying discourse relation markers, we propose the use of syntactic paraphrase methods.

1 Introduction

Discourse relations that hold between the content of clauses and of sentences – including relations of cause, contrast, elaboration, and temporal ordering – are important for natural language processing tasks that require sensitivity to more than just a single sentence, such as summarization, information extraction, and generation. In written text, discourse relations have usually been considered to be signaled either explicitly, as lexicalized with some word or phrase, or implicitly due to adjacency. Thus, while the causal relation between the situations described in the two clauses in Ex. (1) is signalled explicitly by the connective *As a result*, the same relation is conveyed implicitly in Ex. (2).

(1) John was tired. As a result he left early.

(2) John was tired. He left early.

This paper focusses on the problem of how to characterize and identify explicit signals of discourse relations, exemplified in Ex. (1). To refer to all such signals, we use the term "discourse relation markers" (DRMs). Past research (e.g., (Halliday and Hasan, 1976; Martin, 1992; Knott, 1996), among others) has assumed that DRMs are frozen or fixed expressions from a few well-defined syntactic classes, such as conjunctions, adverbs, and prepositional phrases. Thus the literature presents *lists* of DRMs, which researchers try to make as complete as possible for their chosen language. In annotating lexicalized discourse relations of the Penn Discourse Treebank (Prasad et al., 2008), this same assumption drove the initial phase of annotation. A list of "explicit connectives" was collected from various sources and provided to annotators, who then searched for these expressions in the text and annotated them, along with their arguments and senses. The same assumption underlies methods for automatically identifying DRMs (Pitler and Nenkova, 2009). Since expressions functioning as DRMs can also have non-DRM functions, the task is framed as one of classifying given individual tokens as DRM or not DRM.

In this paper, we argue that placing such syntactic and lexical restrictions on DRMs limits a proper understanding of discourse relations, which can be realized in other ways as well. For example, one should recognize that the instantiation (or exemplification) relation between the two sentences in Ex. (3) is explicitly signalled in the second sentence by the phrase *Probably the most egregious example is*, which is sufficient to express the instantiation relation.

(3) Typically, these laws seek to prevent executive branch officials from inquiring into whether certain federal programs make any economic sense or proposing more market-oriented alternatives to regulations. Probably the most egregious example is a

> proviso in the appropriations bill for the executive office that prevents the president's Office of Management and Budget from subjecting agricultural marketing orders to any cost-benefit scrutiny.

Cases such as Ex. (3) show that identifying DRMs cannot simply be a matter of preparing a list of fixed expressions and searching for them in the text. We describe in Section 2 how we identified other ways of expressing discourse relations in the PDTB. In the current version of the corpus (PDTB 2.0.), they are labelled as *AltLex* (*alternative lexicalizations*), and are "discovered" as a result of our lexically driven annotation of discourse relations, including explicit as well as implicit relations. Further analysis of AltLex annotations (Section 3) leads to the thesis that *DRMs are a lexically open-ended class of elements which may or may not belong to well-defined syntactic classes*. The open-ended nature of DRMs is a challenge for their automated identification, and in Section 4, we point to some lessons we have already learned from this annotation. Finally, we suggest that methods used for automatically generating candidate paraphrases may help to expand the set of recognized DRMs for English and for other languages as well (Section 5).

2 AltLex in the PDTB

The Penn Discourse Treebank (Prasad et al., 2008) constitutes the largest available resource of lexically grounded annotations of discourse relations, including both explicit and implicit relations.[1] Discourse relations are assumed to have two and only two arguments, called Arg1 and Arg2. By convention, Arg2 is the argument syntactically associated with the relation, while Arg1 is the other argument. Each discourse relation is also annotated with one of the several senses in the PDTB hierarchical sense classification, as well as the attribution of the relation and its arguments. In this section, we describe how the annotation methodology of the PDTB led to the identification of the AltLex relations.

Since one of the major goals of the annotation was to lexically ground each relation, a first step in the annotation was to identify the explicit markers of discourse relations. Following standard practice, a list of such markers – called "explicit connectives" in the PDTB – was collected from various sources (Halliday and Hasan, 1976; Martin, 1992; Knott, 1996; Forbes-Riley et al., 2006).[2] These were provided to annotators, who then searched for these expressions in the corpus and marked their arguments, senses, and attribution.[3] In the pilot phase of the annotation, we also went through several iterations of updating the list, as and when annotators reported seeing connectives that were not in the current list. Importantly, however, connectives were constrained to come from a few well-defined syntactic classes:

- *Subordinating conjunctions:* e.g., because, although, when, while, since, if, as.
- *Coordinating conjunctions:* e.g., and, but, so, either..or, neither..nor.
- *Prepositional phrases:* e.g., as a result, on the one hand..on the other hand, insofar as, in comparison.
- *adverbs:* e.g., then, however, instead, yet, likewise, subsequently

Ex. (4) illustrates the annotation of an explicit connective. (In all PDTB examples in the paper, Arg2 is indicated in boldface, Arg1 is in italics, the DRM is underlined, and the sense is provided in parentheses at the end of the example.)

(4) *U.S. Trust, a 136-year-old institution that is one of the earliest high-net worth banks in the U.S., has faced intensifying competition from other firms that have established, and heavily promoted, private-banking businesses of their own.* As a result, **U.S. Trust's earnings have been hurt.** (Contingency:Cause:Result)

After all explicit connectives in the list were annotated, the next step was to identify implicit discourse relations. We assumed that such relations are triggered by adjacency, and (because of resource limitations) considered only those that held between sentences within the same paragraph. Annotators were thus instructed to supply a connective – called "implicit connective" – for

[1] http://www.seas.upenn.edu/˜pdtb

[2] All explicit connectives annotated in the PDTB are listed in the PDTB manual (PDTB-Group, 2008).

[3] These guidelines are recorded in the PDTB manual.

each pair of adjacent sentences, *as long as the relation was not already expressed with one of the explicit connectives provided to them*. This procedure led to the annotation of implicit connectives such as *because* in Ex. (5), where a causal relation is inferred but no explicit connective is present in the text to express the relation.

(5) *To compare temperatures over the past 10,000 years, researchers analyzed the changes in concentrations of two forms of oxygen.* (Implicit=because) **These measurements can indicate temperature changes**, . . . (Contingency:Cause:reason)

Annotators soon noticed that in many cases, they were not able to supply an implicit connective. Reasons supplied included (a) "there is a relation between these sentences but I cannot think of a connective to insert between them", (b) "there is a relation between the sentences for which I can think of a connective, but it doesn't sound good", and (c) "there is no relation between the sentences". For all such cases, annotators were instructed to supply "NONE" as the implicit connective. Later, we sub-divided these "NONE" implicits into "EntRel", for the (a) type above (an entity-based coherence relation, since the second sentence seemed to continue the description of some entity mentioned in the first); "NoRel" (no relation) for the (c) type; and "AltLex", for the (b) type, which we turn to next.

Closer investigation of the (b) cases revealed that the awkwardness perceived by annotators when inserting an implicit connective was due to *redundancy in the expression of the relation*: Although no explicit connective was present to relate the two sentences, some other expression appeared to be doing the job. This is indeed what we found. Subsequently, instances of AltLex were annotated if:

1. A discourse relation can be inferred between adjacent sentences.
2. There is no explicit connective present to relate them.
3. The annotator is not able to insert an implicit connective to express the inferred relation (having used "NONE" instead), because inserting it leads to an awkward redundancy in expressing the relation.

Under these conditions, annotators were instructed to look for and mark as *Altlex*, whatever *alternative expression* appeared to denote the relation. Thus, for example, Ex. (6) was annotated as AltLex because although a causal relation is inferred between the sentences, inserting a connective like *because* makes expression of the relation redundant. Here the phrase *One reason is* is taken to denote the relation and is marked as *AltLex*.

(6) *Now, GM appears to be stepping up the pace of its factory consolidation to get in shape for the 1990s.* **One reason is mounting competition from new Japanese car plants in the U.S. that are pouring out more than one million vehicles a year at costs lower than GM can match**. (Contingency:Cause:reason)

The result of this procedure led to the annotation of 624 tokens of AltLex in the PDTB. We turn to our analysis of these expressions in the next section.

3 What is found in AltLex?

Several questions arise when considering the AltLex annotations. What kind of expressions are they? What can we learn from their syntax? Do they project discourse relations of a different sort than connectives? How can they be identified, both during manual annotation and automatically? To address these questions, we examined the AltLex annotation for annotated senses, and for common lexico-syntactic patterns extracted using alignment with the Penn Treebank (Marcus et al., 1993).[4]

3.1 Lexico-syntactic Characterization

We found that we could partition AltLex annotation into three groups by (a) whether or not they belonged to one of the syntactic classes admitted as explicit connectives in the PDTB, and (b) whether the expression was frozen (ie, blocking free substitution, modification or deletion of any of its parts) or open-ended. The three groups are shown in Table 1 and discussed below.

[4] The source texts of the PDTB come from the Penn Treebank (PTB) portion of the Wall Street Journal corpus. The PDTB corpus provides PTB tree alignments of all its text span annotations, including connectives, AltLex's, arguments of relations, and attribution spans.

AltLex Group	No (%)	Examples
Syntactically admitted, lexically frozen	92 (14.7%)	quite the contrary (ADVP), for one thing (PP), as well (ADVP), too (ADVP), soon (ADVP-TMP), eventually (ADVP-TMP), thereafter (RB), even (ADVP), especially (ADVP), actually (ADVP), still (ADVP), only (ADVP), in response (PP)
Syntactically free, lexically frozen	54 (8.7%)	What's more (SBAR-ADV), Never mind that (ADVP-TMP;VB;DT), To begin with (VP), So (ADVP-PRD-TPC), Another (DT), further (JJ), As in (IN;IN), So what if (ADVP;IN), Best of all (NP)
Syntactically and lexically free	478 (76.6%)	That compares with (NP-SBJ;VBD;IN), After these payments (PP-TMP), That would follow (NP-SBJ;MD;VB), The plunge followed (NP-SBJ;VBD), Until then (PP-TMP), The increase was due mainly to (NP-SBJ;VBD;JJ;RB;TO), That is why (NP-SBJ;VBZ;WHADVP), Once triggered (SBAR-TMP)
TOTAL	624	–

Table 1: Breakdown of AltLex by Syntactic and Lexical Flexibility. Examples in the third column are accompanied (in parentheses) with their PTB POS tags and constituent phrase labels obtained from the PDTB-PTB alignment.

Syntactically admitted and lexically frozen: The first row shows that 14.7% of the strings annotated as AltLex belong to syntactic classes admitted as connectives and are similarly frozen. (Syntactic class was obtained from the PDTB-PTB alignment.) So, despite the effort in preparing a list of connectives (cf. Section 1), additional ones were still found in the corpus through AltLex annotation. This suggests that any pre-defined list of connectives should only be used to guide annotators in a strategy for "discovering" connectives.

Syntactically free and lexically frozen: AltLex expressions that were frozen but belonged to syntactic classes other than those admitted for the PDTB explicit connectives accounted for 8.7% (54/624) of the total (Table 1, row 2). For example, the AltLex *What's more* (Ex. 7) is parsed as a clause (SBAR) functioning as an adverb (ADV). It is also frozen, in not undergoing any change (eg, *What's less*, *What's bigger*, etc.[5]

(7) Marketers themselves are partly to blame: *They've increased spending for coupons and other short-term promotions at the expense of image-building advertising.* **What's more, a flood of new products has given consumers a dizzying choice of brands, many of which are virtually carbon copies of one other**. (Expansion:Conjunction)

Many of these AltLex annotations do not constitute a single constituent in the PTB, as with *Never mind that*. These cases suggest that either the restrictions on connectives as frozen expressions should be relaxed to admit all syntactic classes, or the syntactic analyses of these *multi-word expressions* is irrelevant to their function.

Both syntactically and lexically free: This third group (Table 1, row 3) constitutes the majority of AltLex annotations – 76.6% (478/624). Additional examples are shown in Table 2. Common syntactic patterns here include subjects followed by verbs (Table 2a-c), verb phrases with complements (d), adverbial clauses (e), and main clauses with a subordinating conjunction (f).

All these AltLex annotations are freely modifiable, with their fixed and modifiable parts shown in the regular expressions defined for them in Table 2. Each has a fixed "core" phrase shown as lexical tokens in the regular expression, e.g, *consequence of*, *attributed to*, plus obligatory and optional elements shown as syntactic labels. Optional elements are shown in parentheses. <NX> indicates any noun phrase, <PPX>, any prepositional phrase, <VX>, any verb phrase, and

[5] Apparently similar headless relative clauses such as *What's more exciting* differ from *What's more* in not functioning as adverbials, just as NPs.

AltLex String	**AltLex Pattern**
(a) A consequence of their departure could be ...	<DTX> consequence (<PPX>) <VX>
(b) A major reason is ...	<DTX> (<JJX>) reason (<PPX>) <VX>
(c) Mayhap this metaphorical connection made ...	(<ADVX>) <NX> made
(d) ... attributed the increase to ...	attributed <NX> to
(e) Adding to that speculation ...	Adding to <NX>
(f) That may be because ...	<NX> <VX> because

Table 2: Complex AltLex strings and their patterns

<JJX>, any adjectival phrase

These patterns show, for example, that other variants of the identified AltLex *A major reason is* include *The reason is*, *A possible reason for the increase is*, *A reason for why we should consider DRMs as an open class is*, etc. This is robust support for our claim that DRMs should be regarded as an open class: The task of identifying them cannot simply be a matter of checking an *a priori* list.

Note that the optional modification seen here is clearly also possible with many explicit connectives such as *if* (eg, *even if just if*, *only if*), as shown in Appendix C of the PDTB manual (PDTB-Group, 2008). This further supports the thesis that DRMs should be treated as an open class that includes explicit connectives.

3.2 Semantic Characterization

AltLex strings were annotated as denoting the discourse relation that held between otherwise unmarked adjacent utterances (Section 2). We found them to convey this relation in much the same way as anaphoric discourse adverbials. According to (Forbes-Riley et al., 2006), discourse adverbials convey both the discourse relation and an anaphoric reference to its Arg1. The latter may be either explicit (e.g., through the use of a demonstrative like “this” or “that”), or implicit. Thus, both *as a result of that* and *as a result* are discourse adverbials in the same way: the latter refers explicitly to Arg1 via the pronoun “that”, while former does so via an implicit internal argument. (A *result* must be a result of something.)

The examples in Table 2 make this same two–part semantic contribution, albeit with more complex expressions referring to Arg1 and more complex modification of the expression denoting the relation. For example, in the AltLex shown in (Table 2c), *Mayhap this metaphorical connection made* (annotated in Ex. (8)), the relation is denoted by the causal verb *made*, while Arg1 is referenced through the definite description *this metaphorical connection*. In addition, the adverb *Mayhap* further modifies the relational verb.

(8) *Ms. Bartlett’s previous work, which earned her an international reputation in the non-horticultural art world, often took gardens as its nominal subject.* **Mayhap this metaphorical connection made the BPC Fine Arts Committee think she had a literal green thumb**. (Contingency:Cause:Result)

These complex AltLex’s also raise the question of why we find them at all in language. One part of the answer is that these complex AltLex’s are used to convey more than just the meaning of the relation. In most cases, we found that substituting the AltLex with an adverbial connective led to some aspect of the meaning being lost, as in Ex. (9-10). Substituting *For example* for the AltLex with an (necessary) accompanying paraphrase of Arg2 loses the information that the example provided as Arg2 is possibly the most egregious one. The connective *for example* does not allow similar modification. This means that one must use a different strategy such as an AltLex expression.

(9) *Typically, these laws seek to prevent executive branch officials from inquiring into whether certain federal programs make any economic sense or proposing more market-oriented alternatives to regulations.* **Probably the most egregious example is a proviso in the appropriations bill for the executive office that prevents the president’s Office of Management and Budget from subjecting agricultural marketing orders to any cost-benefit scrutiny**. (Expansion:Instantiation)

(10) For example, a proviso in the appropriations bill for the executive office prevents the president’s Of-

fice of Management and Budget from subjecting agricultural marketing orders to any cost-benefit scrutiny.

Another part of the answer to *Why AltLex?* is that it can serve to convey a relation for which the lexicon lacks an adverbial connective. For example, while English has several adverbial connectives that express a "Cause:Consequence" relation (eg, *as a result*, *consequently*, etc.), it lacks an adverbial connective expressing "Cause:Reason" (or explanation) albeit having at least two subordinating conjunctions that do so (*because* and *since*). Thus, we find an AltLex whenever this relation needs to be expressed between sentences, as shown in Ex. (11).

(11) *But a strong level of investor withdrawals is much more unlikely this time around*, fund managers said. **A major reason is that investors already have sharply scaled back their purchases of stock funds since Black Monday.** (Contingency:Cause:reason)

Note, however, that even for such relations such as Cause:Reason, it is still not the case that a list of canned expressions will be sufficient to generate the Altlex or to identify them, since this relation can itself be further modified. In Ex. (12), for example, the writer intends to convey that there are multiple reasons for the walkout, although only one of them is eventually specified in detail.

(12) *In Chile, workers at two copper mines, Los Bronces and El Soldado*, which belong to the Exxon-owned Minera Disputada, *yesterday voted to begin a full strike tomorrow*, an analyst said. **Reasons for the walkout**, the analyst said, **included a number of procedural issues, such as a right to strike**. (Contingency:Cause:reason)

4 Lessons learned from AltLex

Like all lexical phenomena, DRMs appear to have a power-law distribution, with some very few high-frequency instances like (*and*, *but*), a block of mid-frequency instances (eg, *after*, *because*, *however*), and many many low-frequency instances in the "long tail" (eg, *much as*, *on the contrary*, *in short*, etc.). Given the importance of DRMs for recognizing and classifying discourse relations and their arguments, what have we learned from the annotation of AltLex?

First, the number of expressions found through AltLex annotation, that belong to syntactic classes admitted as connectives and also similarly frozen (Table 1, row 1) shows that even in the PDTB, there are additional instances of what we have taken to be explicit connectives. By recognizing them and unambiguously labelling their senses, we will start to reduce the number of "hard cases" of implicit connectives whose sense has to be recognized (Marcu and Echihabi, 2002; Sporleder and Lascarides, 2008; Pitler et al., 2009; Lin et al., 2009). Secondly, the number of tokens of expressions from other syntactic classes that have been annotated as AltLex (Table 1, rows 2 and 3) may actually be higher than was caught via our AltLex annotation, thus making them even more important for discourse processing. To assess this, we selected five of them and looked for all their tokens in the WSJ raw files underlying both the PTB and the PDTB. After eliminating those tokens that had already been annotated, we judged whether the remaining ones were functioning as connectives. Table 3 shows the expressions we used in the first column, with the second and third columns reporting the number of tokens annotated in PDTB, and the number of additional tokens in the WSJ corpus functioning as connectives. (The asterisk next to the expressions is a wild card to allow for variations along the lines discussed for Table 2.) These results show that these DRMs occur two to three times more frequently than already annotated.

Increased frequencies of AltLex occurrence are also observed in discourse annotation projects undertaken subsequent to the PDTB, since they were able to be more sensitive to the presence of AltLex. The Hindi Discourse Relation Bank (HDRB) (Oza et al., 2009), for example, reports that 6.5% of all discourse relations in the HDRB have been annotated as AltLex, compared to 1.5% in the PDTB. This also provides cross-linguistic evidence of the importance of recognizing the full range of DRMs in a language.

5 Identifying DRMs outside the PDTB

As the set of DRMs appears to be both open-ended and distributed like much else in language, with a very long tail, it is likely that many are missing from the one-million word WSJ corpus annotated in the PDTB 2.0. Indeed, in annotating En-

AltLex	**Annotated**	**Unannotated**
The reason*	8	15
That's because	11	16
The result*	12	18
That/This would*	5	16
That means	11	17
TOTAL	47	82

Table 3: Annotated and Unannotated instances of AltLex

glish biomedical articles with discourse relations, Yu et al (2008) report finding many DRMs that don't appear in the WSJ (e.g., *as a consequence*). If one is to fully exploit DRMs in classifying discourse relations, one must be able to identify them all, or at least many more of them than we have to date. One method that seems promising is Callison-Burch's paraphrase generation through back-translation on pairs of word-aligned corpora (Callison-Birch, 2007). This method exploits the frequency with which a word or phrase is back translated (from texts in language A to texts in language B, and then back from texts in language B to texts in language A) across a range of pivot languages, into other words or phrases.

While there are many factors that introduce low-frequency noise into the process, including lexical ambiguity and errors in word alignment, Callison-Burch's method benefits from being able to use the many existing word-aligned translation pairs developed for creating translation models for SMT. Recently, Callison-Burch showed that paraphrase errors could be reduced by syntactically constraining the phrases identified through back-translation to ones with the same syntactic category as assigned to the source (Callison-Birch, 2008), using a large set of syntactic categories similar to those used in CCG (Steedman, 2000).

For DRMs, the idea is to identify through back-translation, instances of DRMs that were neither included in our original set of explicit connective nor subsequently found through AltLex annotation. To allow us to carry out a quick pilot study, Callison-Burch provided us with back-translations of 147 DRMs (primarily explicit connectives annotated in the PDTB 2.0, but also including a few from other syntactic classes found through AltLex annotation). Preliminary analysis of the results reveals many DRMs that don't appear anywhere in the WSJ Corpus (eg, *as a consequence*, *as an example*, *by the same token*), as well as additional DRMs that appear in the corpus but were not annotated as AltLex (e.g., *above all*, *after all*, *despite that*). Many of these latter instances appear in the initial sentence of a paragraph, but the annotation of implicit connectives — which is what led to AltLex annotation in the first place (Section 2) — was not carried out on these sentences.

There are two further things to note before closing this discussion. First, there is an additional source of noise in using back-translation paraphrase to expand the set of identified DRMs. This arises from the fact that discourse relations can be conveyed either explicitly or implicitly, and a translated text may not have made the same choices vis-a-vis explicitation as its source, causing additional word alignment errors (some of which are interesting, but most of which are not). Secondly, this same method should prove useful for languages other English, although there will be an additional problem to overcome for languages (such as Turkish) in which DRMs are conveyed through morphology as well as through distinct words and phrases.

6 Related work

We are not the first to recognize that discourse relations can realized by more than just one or two syntactic classes. Halliday and Hasan (1976) document prepositional phrases like *After that* being used to express conjunctive relations. More importantly, they note that any definite description can be substituted for the demonstrative pronoun.

Similarly, Taboada (2006), in looking at how often RST-based rhetorical relations are realized by discourse markers, starts by considering only adverbials, prepositional phrases, and conjunctions, but then notes the occurrence of a single instance of a nominal fragment *The result* in her corpus. Challenging the RST assumption that the basic unit of a discourse is a clause, with discourse relations holding between adjacent clausal units, Kibble (1999) provides evidence that *informational* discourse relations (as opposed to *intentional* discourse relations) can hold intra-clausally as well, with the relation "verbalized" and its arguments realized as nominalizations, as in *Early treatment with Brand X* can prevent *a cold sore developing*. Since his focus is intra-clausal, he does not observe that verbalized discourse relations can hold across sentences as well, where a verb and one of its arguments function similarly to a discourse adverbial, and in the end, he does not provide a proposal for how to systematically identify these alternative realizations. Le Huong et al. (2003), in developing an algorithm for recognizing discourse relations, consider non-verbal realizations (called NP cues) in addition to verbal realizations (called VP cues). However, they provide only one example of such a cue ("the result"). Like Kibble (1999), Danlos (2006) and Power (2007) also focus only on identifying verbalizations of discourse relations, although they do consider cases where such relations hold across sentences.

What has not been investigated in prior work is the basis for the alternation between connectives and AltLex's, although there are several accounts of why a language may provide more than one connective that conveys the same relation. For example, the alternation in Dutch between *dus* ("so"), *daardoor* ("as a result"), and *daarom* ("that's why") is explained by Pander Maat and Sanders (2000) as having its basis in "subjectivity".

7 Conclusion and Future Work

Categorizing and identifying the range of ways in which discourse relations are realized is important for both discourse understanding and generation. In this paper, we showed that existing practices of cataloguing these ways as lists of closed class expressions is problematic. We drew on our experience in creating the lexically grounded annotations of the Penn Discourse Treebank, and showed that markers of discourse relations should instead be treated as open-class items, with unconstrained syntactic possibilities. Manual annotation and automatic identification practices should develop methods in line with this finding if they aim to exhaustively identify all discourse relation markers.

Acknowledgments

We want to thank Chris Callison-Burch, who graciously provided us with EuroParl back-translation paraphrases for the list of connectives we sent him. This work was partially supported by NSF grant IIS-07-05671.

References

Callison-Birch, Chris. 2007. *Paraphrasing and Translation*. Ph.D. thesis, School of Informatics, University of Edinburgh.

Callison-Birch, Chris. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Danlos, Laurence. 2006. Discourse verbs. In *Proceedings of the 2nd Workshop on Constraints in Discourse*, pages 59–65, Maynooth, Ireland.

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

Halliday, M. A. K. and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Huong, LeThanh, Geetha Abeysinghe, and Christian Huyck. 2003. Using cohesive devices to recognize rhetorical relations in text. In *Proceedings of 4th Computational Linguistics UK Research Colloquium (CLUK 4)*, University of Edinburgh, UK.

Kibble, Rodger. 1999. Nominalisation and rhetorical structure. In *Proceedings of ESSLLI Formal Grammar conference*, Utrecht.

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

Lin, Ziheng, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Singapore.

Maat, Henk Pander and Ted Sanders. 2000. Domains of use or subjectivity? the distribution of three dutch causal connectives explained. *TOPICS IN ENGLISH LINGUISTICS*, pages 57–82.

Marcu, Daniel and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the Association for Computational Linguistics*.

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

Martin, James R. 1992. *English text: System and structure*. Benjamins, Amsterdam.

Oza, Umangi, Rashmi Prasad, Sudheer Kolachina, Dipti Mishra Sharma, and Aravind Joshi. 2009. The hindi discourse relation bank. In *Proceedings of the ACL 2009 Linguistic Annotation Workshop III (LAW-III)*, Singapore.

Pitler, Emily and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the Joint Conference of the 47th Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing*, Singapore.

Pitler, Emily, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing*.

Power, Richard. 2007. Abstract verbs. In *ENLG '07: Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 93–96, Morristown, NJ, USA. Association for Computational Linguistics.

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

PDTB-Group. 2008. The Penn Discourse TreeBank 2.0 Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania.

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

Steedman, Mark. 2000. *The Syntactic Process*. MIT Press, Cambridge MA.

Taboada, Maite. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567–592.

Yu, Hong, Nadya Frid, Susan McRoy, P Simpson, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2008. Exploring discourse connectivity in biomedical text for text mining. In *Proceedings of the 16th Annual International Conference on Intelligent Systems for Molecular Biology BioLINK SIG Meeting*, Toronto, Canada.