Reliable Discourse Markers for Contrast Relations

Jennifer Spenader and Anna Lobanova
Department of Artificial Intelligence
University of Groningen
{j.spenader|a.lobanova}@ai.rug.nl

Abstract

Using the RST annotated corpus [Carlson *et al.*, 2003], we use simple statistics on the distribution of discourse markers or cue phrases as evidence of the three-way distinction of Contrast relations, CONTRAST, ANTITHESIS and CONCESSION, recognized in standard Rhetorical Structure Theory (RST, Mann and Thompson 1987). We also show that *however*, an intuitive marker of Contrast, is not actually used statistically significantly more often in Contrast relations than in Cause-Effect relations. These results highlight the need for empirically based discourse marker identification rather than the intuitive method that is the current norm.

1 Introduction

Contrast is a central rhetorical relation. It is one of the most frequent, as shown by discourse annotation projects. It seems to have a clear, intuitive semantic meaning, and has been argued to interact with other linguistic structures like VP-ellipsis (see e.g. Kehler 2000). Finally, it is instinctively associated with several very clear discourse markers, such as e.g. *however*, *although* and *but*.

However, there is a lack of consensus about whether or not there are qualitatively different Contrast relations: RST (Rhetorical Structure Theory) recognizes three different types: CONTRAST proper, ANTITHESIS and CONCESSION, Wolf and Gibson [2005] recognize two, *denial of expectation* and *contrast*, and Segmented Discourse Representation Theory (SDRT, Asher and Lascarides 2003) recognizes one: Contrast.

In this paper we use the annotated RST corpus [Carlson *et al.*, 2003] and simple lexical cooccurrance statistics to determine if intuitive discourse markers of contrast reliably identify Contrast from Cause-Effect relations

Proceedings of the 8th International Conference on Computational Semantics, pages 210–221,
Tilburg, January 2009.

and if the markers also distinguish between the three-way distinction made in RST. The distribution of markers shows that intuition can be surprisingly wrong, e.g. *however* was not a reliable marker of Contrast. We also found that the different RST Contrast relations can be distinguished by their markers. These results illustrate the need for empirically testing intuitively identified rhetorical relation markers, and argue against collapsing the Contrast distinctions, as has been done in many discourse annotation schemes.

2 Contrast as a rhetorical relation

Theoretically, RST leaves the number of relations recognized up to the annotator [Mann and Thompson, 1987], but in the manually annotated RST corpus [Carlson *et al.*, 2003] 78 relations are stipulated, including three Contrast relations: CONTRAST, ANTITHESIS and CONCESSION. As mentioned above, Wolf and Gibson [2005] recognize two Contrast relations among the 11 relations they distinguish between, 'violated expectation' and 'contrast', a distinction which seems to have been inherited from Hobbs [1985], who may have in turn taken it from Lakoff [1971]. In the manual Reese *et al.* [2007] for the annotation of texts according to SDRT, there is only one contrast relation among the 14 relations recognized. Thus RST recognizes the greatest number of contrast types, but there is no empirical evidence supporting these or any other distinctions.

The distinctions between different types of Contrast found in current discourse annotation schemes seem to have been adapted from theoretical linguistic work on contrast that sought to characterize the way in which the conjunction *but* differs from *and*. Lakoff [1971] made a distinction between what she called *denial of expectation* contrast and *semantic opposition* uses of *but*, e.g.

(1) It's raining but I'm taking an umbrella.

(2) John is tall but Bill is short. (Lakoff 1971: 133)

Denial of expectation has semantically been interpreted as a case where the first conjunct implicates a proposition that the second conjunct denies, e.g. in (1) "It's raining" implicates the speaker will get wet, while having an umbrella negates this implication. *Semantic opposition* contrast on the other hand is characterized by the fact that the conjuncts have parallel elements contrasted along one dimension, e.g. in (2), *John* and *Bill* are humans contrasted according to their height.

The three RST relations seem to preserve the same the distinction. ANTITHESIS and CONTRAST are described as having contrast "happen in only one or few respects, while everything else can remain the same in other respects." (Annotation manual, [Carlson and Marcu, 2001] same wording in both definitions.), clearly echoing the definition of *semantic opposition*: ANTITHESIS and CONTRAST only differ in terms of symmetry, realized in terms of nuclearity in RST. In multinuclear CONTRAST neither of the conjuncts should be more prominent or more connected with the rest of the discourse than the other, but in a mononuclear ANTITHESIS relation the nucleus will be more prominent.

ANTITHESIS 1 [Although Exxon spent heavily during the latest quarter to clean up the Alaskan shoreline blackened by its huge oil spill,]^{1A} [those expenses as well as the cost of a continuing spill-related program are covered by $880 million in charges taken during the first half.]^{1B} (wsj$_{1311}$)

ANTITHESIS 2 [A hearing is set for Nov. 15,]^{2A} [but participants don't expect a resolution until July 1990.]^{2B} (wsj$_{1145}$)

CONTRAST 3 [Import values are calculated on a cost, insurance and freight (c.i.f.) basis,]^{3A} [while exports are accounted for on a free-on-board (f.o.b.) basis.]^{3B} (wsj$_{0615}$)

CONTRAST 4 [The clash of ideologies survives this treatment,]^{4A} [but the nuance and richness of Gorky's individual characters have vanished in the scuffle.]^{4B} (wsj$_{0615}$)

For a CONCESSION relation the contrast is argued to be the result of an unexpected situation, and the definition even says it involves a *denial of expectation*.

> "The situation indicated in the nucleus is contrary to expectation in the light of the information presented in the satellite. In other words, a CONCESSION relation is always characterized by a violated expectation. (Compare to ANTITHESIS.) In some cases, which text span is the satellite and which is the nucleus do not depend on the semantics of the spans, but rather on the intention of the writer." (Annotation manual, Carlson and Marcu [2001])

Examining two examples from the corpus below what we can see is that we should not have the kind of parallel elements typical of CONTRAST and ANTITHESIS.

CONCESSION 5 [Its 1,400-member brokerage operation reported an estimated $5 million loss last year,][5A] [although Kidder expects it to turn a profit this year.][5B] (wsj$_{0604}$)

CONCESSION 6 [While there have been no reports of similar sudden unexplained deaths among diabetics in the U.S.,][6A] [Dr. Sobel said the FDA plans to examine Dr. Toseland's evidence and is considering its own study here.][6B] (wsj$_{0690}$)

However, these two categories are hard to apply straightforwardly to many examples. Further, numerous linguistic papers (e.g. Foolen 1991, Winter and Rimon 1994 and Spenader and Stulp 2007)have argued that the distinction between *denial of expectation* and *semantic opposition* is artificial, and that to correctly interpret a sentence such as (2) in a discourse it is necessary to have a context such as e.g. "All Dutch people aren't giants", the interpretation becomes the same as for a *denial of expectation*.

Just how easy is it to distinguish an ANTITHESIS relation from CONTRAST or CONCESSION? Carlson *et al.* [2003] present kappa scores for subsets of the corpus ranging from 0.6 to 0.79 for the set of 78 relations, and scores up to 0.82 for a simpler annotation scheme where the 78 categories were collapsed into 16 supersets, including one Contrast set. But they don't report scores for the entire corpus or for sets of particular relations in isolation, so all we can do is evaluate individual examples. CONTRAST 3 and CONTRAST 4 do seem to display parallel elements but what about ANTITHESIS 2. Why isn't the fact that it will take so long to reach a verdict considered a kind of *denial of expectation*? Are the dates the parallel elements in ANTITHESIS 2? The annotation doesn't require explicitly identifying these structures but the definitions imply they should be present. In many ways, ANTITHESIS 2 seems to share more with CONCESSION 6. For CONCESSION 5 we could also easily argue that the brokerage operation and Kidder are parallel elements while profits or losses is the measure of comparison.

In the end, the corpus has a similar problem to all materials with annotations where there is no clear, objective method of categorization. We have to simply accept the annotation as reliable and see if the results we obtain with it makes sense.

3 Previous research analyzing cue words

Taboada [2006] used the RST corpus and a corpus of task oriented dialogues that she annotated with RST relations to identify the most frequently used

discourse markers for a number of RST relations. Most relevant for the current work are her results for unembedded CONCESSION relations. In the RST corpus she found that 90.35% of the relations were marked with some recognisable discourse marker, with the words *but* and *although* contributing to 50% of the marked relations. Other markers she identified were *though*, *despite*, *while*, *even though*, *however*, *still*, *even if*, *even when*, *even yet*, *whether* and *whereas*. Another relevant result concerns the distribution of discourse markers across nuclei and satellites. She found that for CONCESSION, the markers were equally likely to occur in the nucleus or satellite.

The main problem with this study is that it relies on intuition for the initial identification of the CONCESSION markers, and then the frequency with which they intuitively seem to be signaling contrast is used as evidence of the correctness of the initial intuition. But this means that relevant markers might be missed. An even greater problem is that the method does not insure that identified markers are actually characteristic of the relation; they might very well occur just as frequently in other relations. The frequency with which *but* and *although* occur in the CONCESSION relations and intuition makes a strong case for considering these markers of CONCESSION, but markers like e.g. *while*, might be just as likely to occur with a RESULT or a CAUSE relation.

Marcu and Echihabi [2002] used machine learning to develop an automatic classifier for a number of super categories of discourse relations, including Contrast, Cause-Explanation-Evidence, Condition and Elaboration. First, they made a set of patterns based on intuitively identified discourse markers for each discourse relation. They then used these markers to automatically extracted large numbers of examples from two corpora totally more than 42 million English sentences. For example, sentences with a sentence-initial *but* were considered Contrast examples, and sentences with *because* as Cause-Explanation-Evidence. For training, all discourse markers were removed and the stripped sentences were used to train a family of Naïve Bayes classifiers. One reported results was that the classifier that distinguished between CAUSE-EXPLANATION-EVIDENCE and CONTRAST had an accuracy of 87.1%. The level of accuracy is impressive, and surely supports the authors' claim that automatical extraction is a reliable method for finding large number of examples of certain coherence relations. On the other hand, it is not clear what could be achieved, and making more fine-grained distinctions might require less noisy data. For such investigations, intuitively identified discourse markers might not be reliable enough.

Sporleder and Lascarides [2008] compared the performance of rhetorical relation classifiers trained on data with marked and unmarked discourse re-

lations. They chose a subset of five discourse relations including CONTRAST, RESULT, SUMMARY, EXPLANATION and CONTINUATION and a total of 55 discourse markers that according to them unambiguously indicated each of the relations. For example, *but*, *although*, *however*, *whereas* and *yet* were considered to be unambiguous markers of CONTRAST because following SDRT definition Sporleder and Lascarides [2008] assumed that there is only one type of Contrast. The choice of discourse markers was based on Oates [2000] and authors' introspection of randomly extracted examples. What is relevant to our research is that both studies (Marcu and Echihabi 2002 and Sporleder and Lascarides 2008) extracted explicitly marked rhetorical relations using a set of discourse markers selected by intuition, without any empirical evidence that the markers are reliable. In addition, no fine-grained distinctions between types of relations (e.g. CONCESSION vs ANTITHESIS) were made.

One way to determine if discourse markers are reliable indicators of the relations we assume they mark is to see if the qualitative difference postulated between the Contrast relations seems to manifest in a distributional difference in the discourse markers used in Contrast relations. An immediate potential objection to this methods is the fact that the RST corpus annotation manual lists a number of intuitively identified discourse markers as potential indicators for many of the relations, including the Contrast relations. For example, it says that the discourse markers *although* and *despite* are discourse markers for CONCESSION and ANTITHESIS, while *however* is a discourse marker for ANTITHESIS and CONTRAST.

Indeed, if we only find evidence that these markers pattern with the mentioned relations, then we cannot determine if this is because the relations themselves are best marked with these markers, or if the annotators were simply influenced by the manual. If, however, we do find some other consistent pattern of discourse markers correlating with each of the Contrast relations, then this would be evidence that these qualitative distinctions are real, rather than merely stipulated by the coding scheme.

A final note, RST allows relations to be embedded in other relations, a feature that seems to be unique to RST, and the RST corpus among other discourse annotated corpora. We think it is important to look both at simple relations and at embedded relations, but in this we depart from much of the earlier work done on studying discourse markers. This has a disadvantage in that it can inflate the counts, because a discourse marker inside a Contrast relation that is in turn embedded inside another Contrast relation will be counted twice as marking Contrast relations. On the other hand, there is no other way to count discourse markers and still take embedded contexts into account.

		Contrast set (%)	Cause-Effect set (%)
p	**though**	**0.11**	0.05
	even	**0.18**	0.11
	despite	**0.06**	0.02
p	**although**	**0.10**	0.02
p	**but**	**0.85**	0.44
	however	0.09	0.06
p	**still**	**0.14**	0.05
p	**while**	**0.15**	0.09
	only	0.15	0.17
	too	**0.08**	0.04

Table 1: Set of three Contrast relations compared with three Cause-Effect relations. Words in bold occur significantly more often in one relation than the other to the degree of $p \leq 0.05$. When a 'p' precedes the word $p \leq 0.009$. The relation in which the word occurred significantly more frequently in has the percent marked in bold. Thus *though* occurred 84 times in the three Contrast relations. The three relations had 75,552 words, so *though* occurred with a frequency of 0.0011, or made up 0.11% of the total words. All tables present the data according to this pattern.

4 Experiments

We used the annotated RST corpus as data [Carlson *et al.*, 2003]. This corpus has approximately 176,000 words composed of 385 articles from the Wall Street Journal portion of the Penn Treebank. We extracted all CONTRAST, ANTITHESIS, CONCESSION, EVIDENCE, CAUSE and RESULT relations,[1] including relations that contained embedded relations. We then use χ^2 tests to check for statistically significant correlations between lexical items and the different coherence relations. We only report results for a small set of closed class words that are particularly likely to be discourse markers.

First, from the results in Table 1 we can see that many terms considered to be typical markers of Contrast do in fact distinguish Contrast relations from Cause-Effect relations. A somewhat surprising result is that *however*, stereotypically considered a marker of contrast, is not used significantly more often in Contrast than in Cause-Effect relations. Also, a number of lexical items that are not generally recognized as discourse markers but which do tend to contribute to Contrast are in fact significant. These include *even*, *still* and the parallel marker *too*.

[1] CONCESSION was the smallest relation, with 15,346 words. CONTRAST was the largest with 35,859. ANTITHESIS relations contained 24,347 words.

		Antithesis+ Concesssion (%)	Contrast (%)
p	**though**	**0.16**	0.05
p	**even**	**0.22**	0.13
p	**despite**	**0.08**	0.03
p	**although**	**0.15**	0.04
	but	0.85	0.44
	however	0.09	0.06
	still	**0.17**	0.11
	while	**0.18**	0.12
	only	0.17	0.13
	too	0.09	0.04

Table 2: Nuclearity compared: Mononuclear ANTITHESIS and CONCESSION compared with multinuclear CONTRAST.

		Antithesis+ Contrast (%)	Concession (%)
p	**though**	0.07	**0.27**
p	**even**	0.14	**0.33**
p	**despite**	0.04	**0.13**
p	**although**	0.08	**0.17**
	but	**0.89**	0.44
	however	0.09	0.08
	still	0.13	0.18
	while	0.16	0.14
	only	0.14	**0.22**
	too	0.08	0.06

Table 3: Contrast types compared: ANTITHESIS and CONTRAST versus CONCESSION.

Next, we examined different groupings of the contrast relations to see if there is evidence that the three categories of contrast distinguished by RST actually show a different distribution of discourse markers.

The three Contrast relations can be further grouped along two features, their nuclearity and the way in which they create the contrastive meaning. ANTITHESIS and CONCESSION are both mononuclear relations while CONTRAST is multinuclear. Are either of these features reflected in the type of discourse markers the relations cooccur with? It is highly possible that nuclearity would limit which discourse markers cooccur with which relations given that nuclearity to a certain degree correlates with the coordinating and subordinating conjunction distinction. To test this question we compared ANTITHESIS and CONCESSION to CONTRAST. The results are in Table 2.

The first thing to notice is that *but* and *too* are no longer significant: they mark ANTITHESIS and CONCESSION equally as well as they mark CONTRAST. We also see that a number of markers that were useful for distinguishing Contrast from Cause-Effect relations are also useful for distinguishing ANTITHESIS and CONCESSION from CONTRAST, occurring significantly more often in ANTITHESIS and CONCESSION, i.e. *though*, *although*, *despite*, *even*, *still* and *while*.

What if we instead group the three relations by the way in which they seem to establish contrast? Remember, from the definitions CONCESSION has to do with a violated expectation between the two discourse units, whereas both CONTRAST and ANTITHESIS should be characterized by a comparison along 'one or more respects'.

		Antithesis (%)	Concession (%)
p	**though**	0.10	**0.27**
p	**even**	0.15	**0.33**
p	**despite**	0.05	**0.13**
	although	0.14	0.17
	but	**0.93**	0.72
	however	0.12	0.08
	still	0.16	0.18
	while	0.20	0.15
	only	0.14	0.22
	too	0.10	0.06

Table 4: ANTITHESIS compared with CONCESSION

		Contrast (%)	Concession (%)
p	**though**	0.05	**0.27**
p	**even**	0.13	**0.33**
p	**despite**	0.03	**0.13**
p	**although**	0.04	**0.17**
	but	0.86	0.72
	however	0.08	0.08
	still	0.11	**0.18**
	while	0.12	0.15
	only	0.13	**0.22**
	too	0.07	0.06

Table 5: CONTRAST compared with CONCESSION

The results in Table 3 show that CONCESSION can be distinguished from ANTITHESIS and CONTRAST by the typical markers *though*, *although*, *even* and *despite*, as well as *only*. The markers *while* and *still* are no longer significant. These results, combined with the results above seem to suggest that CONCESSION is quite different from ANTITHESIS and CONTRAST. Probably these markers are actually just markers of CONCESSION. We can check this by comparing CONCESSION with ANTITHESIS (Table 4) and CONCESSION with CONTRAST (Table 5). What we then see is that *though*, *even* and *despite* distinguish CONCESSION from ANTITHESIS and CONCESSION from CONTRAST. Table 5 shows that *although* also distinguishes CONCESSION from CONTRAST but because this cue does not distinguish ANTITHESIS from CONCESSION we can guess that it is equally as characteristic of ANTITHESIS as it is of CONCESSION. This also explains why it was significantly different from CONTRAST when we collapsed ANTITHESIS with CONCESSION. The same holds for *still*.

5 Discussion and Conclusions

Our first conclusion is that we seem to have found that each relation has a distinctive discourse marker profile and that these results support the three-way distinction, that otherwise seems to be stipulated. Further, in terms of the discourse markers that distinguish them it seems that CONCESSION is much more different from ANTITHESIS and CONTRAST, suggesting that the way in which the contrast relation is established is more relevant to

		Antithesis (%)	Contrast (%)
	though	**0.10**	0.05
	even	0.15	0.13
	despite	0.05	0.03
p	**although**	**0.14**	0.04
	but	0.93	0.86
	however	0.12	0.08
	still	0.16	0.11
	while	0.20	0.12
	only	0.14	0.13
	too	0.10	0.07

Table 6: ANTITHESIS compared with CONTRAST

lexical marking choices than nuclearity and/or symmetry. This is also in line with the two way distinction in the theoretical linguistics. Note also that the reliable discourse markers differ from those suggested in the annotation manual: *although* and *despite* are only reliable markers of CONCESSION, not ANTITHESIS, and *however* doesn't characterize CONTRAST relations at all.

A second result is that by using χ^2 statistics to identify discourse markers we have a reliable and fairly automatic alternative method to the intuitive identification of markers made by much of the existing research. This method can be applied to other discourse relations and may find some surprising results, such as e.g. our finding that *however* is not a reliable unambiguous marker of Contrast when compared with Cause-Effect relations. Of course, it is entirely possible that *however* is a good indicator of Contrast when distinguishing Contrast from e.g. NARRATION. Ideally, we should compare all combinations to derive an exhaustive and data derived list of reliable discourse markers for all relations, but we limit our discussion to a small set of lexical items and only compare Contrast with Cause-Effect relations because of time and space constraints, but this is an obvious next step in our inquiry.

Our results have implications for data oriented approaches using intuition to identify markers to extract examples of coherence relations. Marcu and Echihabi [2002] for example relied solely on discourse markers to extract training data, necessarily so because the method they used requires more data than could feasibly be manually annotated. But our results show that careful testing of the reliability of the discourse markers could improve the quality of the extracted relations. Further, the number of Contrast relations

recognized has to be carefully considered. Treating all Contrast relations as one supercategory, collapsing the RST distinctions as Marcu and Echihabi [2002] and many others have done, may lead to worse results than retaining the distinctions; we know from part of speech tagging for example, that while too many distinctions may make tagging harder, too few can do the same. The results also show that even a modest amount of annotated data can be useful for improving extracted data.

Finally, one of the most obvious problems with all the studies (including this one) on automatically identifying discourse relations is that they only work with marked discourse relations. Our results won't help much in identifying unmarked Contrast relations, yet these relations are very frequent. Carlson *et al.* [2003] have shown that in the corpus of Rhetorical Structure trees only 61 out of 238 contrast relations were marked by a discourse marker. This means that contrastive markers would help to identify only 25% of contrast relations in that corpus. Similarly, Taboada [2006] looked at the RST corpus and a task-oriented dialogue corpus and concluded that most of the relations (between 60-70%) were not signaled by any discourse markers. Finding a solution to these problems will be a challenge for future work.

Acknowledgements

We would like to thank Axel Brink for help with data extraction, and the anonymous reviewers for useful comments. Jennifer Spenader's work was supported by grant 016.064.062 from the Netherlands Organisation for Scientific Research (NWO).

References

Nicholas Asher and Alex Lascarides. *Logics of Conversation.* Cambridge University Press, 2003.

Lynn Carlson and Daniel Marcu. Discourse tagging manual. Technical report, ISI Tech Report ISI-TR-545, July 2001.

L. Carlson, Daniel Marcu, and M.E Okurowski. *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory.* 2003.

Ad Foolen. Polyfunctionality and the semantics of adversative conjunctions. *Multilingual*, 10(1/2):70–92, 1991.

Jerry Hobbs. On the coherence and structure of discourse. Technical report, Report No. CSLI-85-37, Center for the Study of Language and Information, 1985.

Andrew Kehler. Coherence and the resolution of ellipsis. *Linguistics and Philosophy*, 23(6):533–575, 2000.

Robyn Lakoff. *If, ands and buts about conjunction*, chapter Studies in Linguistic Semantics. Holt, Reinhart and Winston, 1971.

William C. Mann and Sandra A. Thompson. *Rhetorical Structure Theory: A theory of text organization.* Information Sciences Institute, Marina del Rey, CA, 1987.

Daniel Marcu and A Echihabi. An unsupervised approach to recognizing discourse relations. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-). Philadelphia, PA, July 7-12, 2002.

Sarah Louise Oates. Multiple discourse marker occurrence: Creating hierarchies for natural language generation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 41–45, 2000.

Brian Reese, Julia Hunter, Nicholas Asher, Pascal Deni, and Jason Baldridge. Reference manual for the analysis and annotation of rhetorical structure (version 1.0), 2007.

Jennifer Spenader and Gert Stulp. Antonymy and contrast relations. In *Seventh International Workshop on Computational Semantics*, Tilburg, 10-11 January 2007.

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

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

Y. Winter and M. Rimon. Contrast and implication in natural language. *Journal of Semantics*, 1994.

Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–288, 2005.