

Topics in Natural Language Processing

Essay

**Discourse Relations:  
a tool to infer the rhetorical structure of  
discourse**

Alessandra Cervone, s1466724

# Contents

<b>Contents</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Discourse Relations as a successful approach to discourse analysis . . . . .	1
1.2 Methodologies . . . . .	6
1.3 Outline of the essay . . . . .	9
<b>2 Computational approaches to discourse analysis using coherence relations</b>	<b>10</b>
2.1 Resources . . . . .	10
2.2 Main approaches in the literature: NLU for written texts . . . . .	12
2.2.1 Discourse relations assignment and discourse parsing . . . . .	12
2.2.2 Applications . . . . .	18
2.3 A new possible direction of research for DRs: Speech processing . . . . .	19
<b>3 Conclusions</b>	<b>22</b>
3.1 Summary . . . . .	22
3.2 Final remarks and future work . . . . .	23

# Chapter 1

## Introduction

### 1.1 Discourse Relations as a successful approach to discourse analysis

- (1) “ ‘If there’s no meaning in it,’ said the King, ‘that saves a world of trouble, you know, as we needn’t try to find any. And yet I don’t know,’ he went on, spreading out the verses on his knee, and looking at them with one eye; ‘I seem to see some meaning in them, after all.’”

Lewis Carroll, *Alice’s Adventures in Wonderland*

Defining the quality that gives a text its meaning is a rather difficult task. Reconstructing the sense of a text involves considering all the elements coming from the lower levels of language, such as the semantics, syntax, morphology of the different sentences in it and their relationships and reconnecting them into a uniform message that the speaker wants to communicate with its discourse. Being the most abstract level of language, discourse is also the most arduous to capture, because various combinations of all these variables can lead to different interpretations of the sense of a text. As in Example 1 where the King, after having defined the process of finding the meaning of a discourse as “a world of trouble”, argues that he maybe could see a sense, a meaningful interpretation in the text they are talking about. The process of assessing the meaning of a text could thus be difficult also for humans, and to a certian extent also subjective.

In linguistics literature the property which gives meaning to a text is defined as its *coherence* (De Beaugrande and Dressler, 1981). This quality is what makes a text a text, rather than a set of casual sentences assembled together. A discourse can be regarded as coherent if we can recognise in it a conceptual unity. More formal definitions of coherence are however rather difficult to formulate, given the abstract nature of this concept. In

order to better understand the idea of coherence the linguistics literature has devised the concept of *cohesion* (Halliday and Hasan, 1976), that is the property that makes the different sentences in a text connected to each other by a series of surface devices, such as referencing entities previously introduced or having a coherent *consecutio temporum* (concordance of the grammatical tenses used in the sentences). In order to be coherent, a text must be cohesive, but cohesion is not a sufficient condition to define a discourse coherent.

- (2) “ ‘Contrariwise,’ continued Tweedledee, ‘if it was so, it might be; and if it were so, it would be; but as it isn’t, it ain’t. That’s logic.’ ”

Lewis Carroll, *Through the Looking-Glass, and What Alice Found There*

Example 2 can be regarded as cohesive, since the different clauses are connected by the reference to the same entity (in the text “it”), but it is nonetheless not coherent, because it does not have a sense. Coherence is thus produced not only by a series of surface features that give continuity to the text, but also, and more importantly by its logical structure.

Furthermore, the elements introduced in the current *discourse model* (following Webber (1978)), which could be the key to understand its logical structure, could be part of a shared world knowledge.

- (3) “ Alice looked round her in great surprise. ‘Why, I do believe we’ve been under this tree the whole time! Everything’s just as it was!’

‘Of course it is,’ said the Queen, ‘what would you have it?’

‘Well, in our country,’ said Alice, still panting a little, ‘you’d generally get to somewhere else—if you ran very fast for a long time, as we’ve been doing.’

‘A slow sort of country!’ said the Queen, ‘Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!’ ”

Lewis Carroll, *Through the Looking-Glass, and What Alice Found There*

In Example 3 there is a misunderstanding between Alice and the Queen because they do not share the same world knowledge. The Queen does not understand why Alice is surprised of the fact that after having run they are in the same place, because in her world the things work differently.

In order to perform the task of correctly assessing the coherence of a text as well as a human in an automatic way we would thus need to take into account all these levels and have a deep understanding of the discourse. This problem can therefore be regarded

as AI complete, since to solve it we would need a system able to have a complete understanding of the text.

Given the difficulty of this problem, various theories have been formulated in the literature to address different subproblems of the automatic analysis of discourse, but no one has attempted so far, to the best of our knowledge, to build a system able to reconstruct the entire structure of discourse connecting all the insights provided by the different approaches together. The lack of such an effort in the direction of a comprehensive framework is due, besides to the difficulty of the task, also to the fact that real interest in computational approaches to discourse analysis started to develop in the literature only in the late 70s with studies such as Webber (1978), Webber (1983), Grosz et al. (1983), Hobbs (1978), Hobbs (1979), Joshi and Kuhn (1979), which provided some of the insights the field is relying upon still now. Another reason why there haven't been attempts to try to connect the different approaches to computational discourse into a common framework is due to the fact that the more abstract nature of the field (compared to other levels of language such as morphology and syntax) has not allowed to create a set of basic units (like phrases in syntax or morphemes in morphology) on which the scientific community agrees. Different approaches to computational discourse have thus defined their own sets of basic units, according to the subproblem they were aiming to solve (ex. Discourse Segments in the intentional approach (Grosz and Sidner, 1986) are defined by a given intention while in approaches aimed at analyzing the topical structure a discourse segment is identified by the orientation towards a given topic (Kan et al., 1998)).

The main research directions in automatic analysis of discourse have concentrated so far mainly on: reconstructing the *intentional* structure of the text, a group of approaches which study the different purposes that the speaker wants the various units to serve (the most famous formalization being the intentional structure theory formulated in Grosz and Sidner (1986)); or its *topical* structure, where the shifts to different topics define the boundaries between the segments of the discourse using for example entity-chains (as in the centering theory (Grosz et al., 1995)); or its *rhetorical* structure, a category of approaches which concentrates on the logical relationships, known as discourse relations, which connect the different sentences or clauses in the text.

In this essay, we will focus on *discourse* relations (DRs, henceforth), also known in the literature as *coherence* or *rhetorical* relations (these different formulations will be used interchangeably), that is on the rhetorical structure of discourse. We have decided to investigate this research direction rather than the other ones mentioned, because it is the one that has proved most successful for computational applications so far. Although the intentional approach seems also very interesting, due to the difficulties in

modelling human intentions this framework has been studied mainly for the dialogues genre. Furthermore, according to our opinion, discourse relations can offer a deeper insight to the real underlying coherence structure of text, compared for example to the topical structure which deals with more surface phenomena. The logical structure is indeed the most important one, the *sine qua non* condition that makes a text coherent and that is thus required for a true understanding of the meaning of the discourse. For these reasons in this essay we have chosen to concentrate on providing an overview of computational approaches which use discourse relations as a heuristic tool to infer the rhetorical structure of texts, rather than on approaches analyzing other types of structure.

Given the fact that many different sets of coherence relations have been proposed during the years in the literature, such as for example in Hobbs (1979), Mann and Thompson (1987), Asher and Lascarides (2003) among the others, there is no shared agreement regarding how can such relations be formally defined. We can however affirm that with the term discourse relation we refer to the logical connection which links together two sentences or clauses in a text. This connection can either be explicitly marked through a discourse connective (such as “but” for the CONTRAST relation) or not marked (when no explicit cue is present in the text to signal the type of relation between two utterances). To better understand the difference between these two types (explicit and implicit) of discourse relation, we can consider the following example:

- (4) “ ‘When we were little,’ the Mock Turtle went on at last, more calmly, though still sobbing a little now and then, ‘we went to school in the sea. The master was an old Turtle — we used to call him Tortoise —’  
‘Why did you call him Tortoise, if he wasn’t one?’ Alice asked.  
‘We called him Tortoise because he taught us,’ said the Mock Turtle angrily: ‘really you are very dull!’  
‘You ought to be ashamed of yourself for asking such a simple question,’ added the Gryphon.”

Lewis Carroll, *Alice’s Adventures in Wonderland*

In Example 4 the characters argue about the discourse relation holding between the clauses “The master was an old Turtle” and “we used to call him Tortoise”. Alice infers that there is an implicit CAUSAL relation between the fact that the master is a turtle and the fact that he was called Tortoise. However, as often happens in Wonderland, the most common inferences regarding language are defeated. In this case, Alice is considered “dull” by the other characters because she did not understand that there was indeed a CAUSAL implicit relation but between the fact that the turtle was a

master and his nickname, thus not between the fact that the master was a turtle and the second clause (as Alice thought). This example shows also how it is a much more subtle task to identify implicit relations compared to explicit ones (as we will see when considering the DRs extraction systems built in the literature).

As previously mentioned, many different sets of coherence relations have been suggested in the literature, connected to various proposals regarding the theoretical frameworks they should be studied in. Many theories proposed that discourse relations in the texts can be composed into a large hierarchical structure which represents the whole discourse as a unit.

In the Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), (Mann and Thompson, 1988), for example, discourse relations are considered to be composed by a nucleus (the most important clause to the purpose of the discourse which can be considered independently) and a satellite (the clause dependent by the nucleus and less important for the aims of the speaker). According to this basic structure a nucleus could then become the satellite of another relation, thus composing a hierarchical rhetorical structure for the entire discourse, as in Figure 1.1, where the nucleus are indicated by a vertical line and their depending satellites are connected to them by a directed arrow.

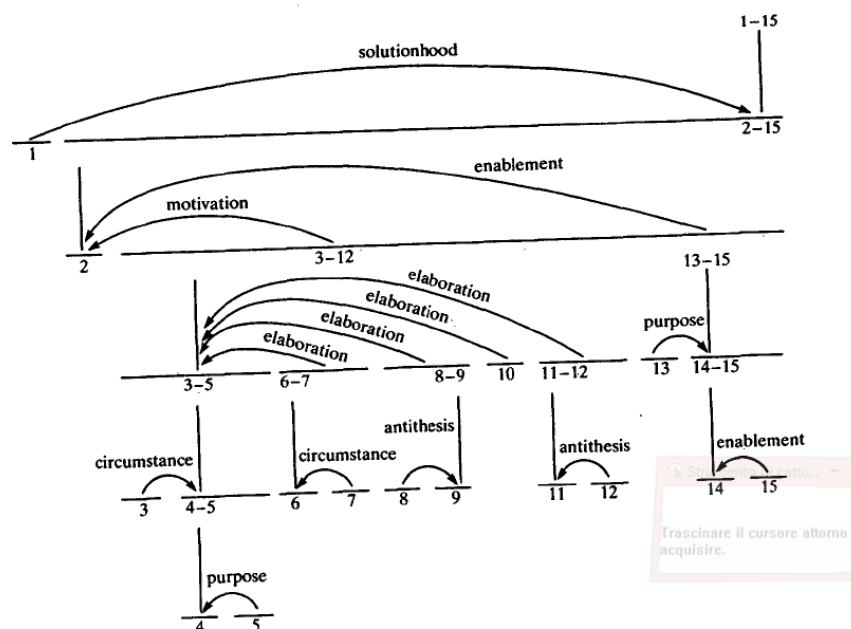


FIGURE 1.1: Example of the hierarchical discourse structure for a whole text in RST from Mann and Thompson (1988).

Mann and Thompson (1988) proposed a set of 110 distinct rhetorical relations, but remarked that this set was not rigid, but rather flexible and open-ended.

Another very well-known approach which proposed its own set of discourse relations and which allows to reconstruct a hierarchical parse for a whole text is the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003), which uses a dynamic and compositional semantic approach to discourse interpretation.

Since the field is so fragmented regarding the number and types of rhetorical relations (apparently a total of 350 have been suggested by various frameworks according to Hovy (1990) ), we are not going to discuss the different proposals of DRs sets under a theoretical perspective, posited that there is a shallow agreement at least on some basic, very common ones, such as CAUSE (where one clause is assumed to be the reason of another one) or CONTRAST (the relation holding between a clause and another clause expressing an idea or an event in contrast with the previous one), which allows the researchers coming from different frameworks to compare at least some of the results.

Our focus will rather be on how coherence relations have been used as a tool to automatically analyze the rhetorical structure of discourse, thus taking an applications perspective. Before doing so, however, we offer a brief introduction to the main methodological tools employed by the approaches dealing with discourse relations (that will be discussed in the next chapter), followed by a more detailed outline of the contents of this essay.

## 1.2 Methodologies

We have already introduced the diversity of the approaches to automatic discourse relations analysis from a theoretical perspective in the previous section. This fragmentation of the field is also mirrored in the methodological tools chosen by various research groups to deal with the challenge of reconstructing the rhetorical structure of a text automatically.

It must be noticed, however, that most of the methods employed for this task belong to the category of machine learning algorithms. This is not surprising since, as previously noticed, the analysis of discourse, being the most abstract level of language, has to take into account all the features coming from the lower levels. Furthermore, the abstract nature of the field makes discourse phenomena often not directly associated to given fixed structure (such as verb agreement for syntax), thus not easily detectable through a set of rules. Machine learning methods, which learn patterns from large amount of training data and which could take into account many different features, are thus the main preferred approach to tackle discourse in general. In particular, being rhetorical relations a manifold phenomenon which could be rather subtle to detect (especially if we consider



implicit DRs), it is quite hard to find ways to successfully use unsupervised approaches (as we will see in 2.2.1). This leads to the field being dominated by supervised methods which treat discourse relations automatic identification as a classification problem, with a few exceptions.

Another type of methodological tools rather often used in computational discourse in general are similarity metrics, generally used to return a measure of the degree of closeness between two objects, which generally in NLP are words. These methods are used in particular in distributional approaches to semantics and discourse.

Finally, sometimes other mathematical models can be used, for example to formalize an optimization problem in order to return the best candidate out of a set of possible choices for a given implementation (given the ambiguity of language these type of problems which require a search through some possibilities are quite common).

### **Machine learning methods**

*Naive Bayes*: a rather simple supervised classification method which however assumes the statistical independence of the predictors (an assumption rarely met in NLP tasks). If we are able to make this assumption for our dataset this algorithm is a very good choice because is very fast to train given its simple architecture (is trained to maximize the likelihood of the training data). To perform well naive bayes requires a large training dataset, but this is not a huge problem thanks to the development of a resource like the PDTB.

*Logistic regression*: a supervised binary classifier which is trained to maximise the conditional log-likelihood of the data in order to estimate the model's parameters. The output of this model is a confidence interval, that is the odds that a given element belongs to a given class (thus a probability estimation).

*Multinomial logistic regression (or MaxEnt)*: a supervised method very often used in computational approaches to discourse relations extraction. It generalizes logistic regression to predict data which could belong to multiple classes (not only two). It is often used instead of naive bayes when we can not assume statistical independence of the predictors (features), since instead of maximizing the likelihood maximizes the conditional likelihood of the data. For the same reason, it is however less fast than naive bayes for large training sets since it requires multiple iterations on the data.

*K-nearest neighbors (kNN)*: a rather simple supervised method which can be used for both classification and regression. It classifies an element according to the characteristic of the elements near (the neighbours) to the one under consideration.

*Support Vector Machines (SVM)*: a supervised binary linear classifier which employs the 'kernel trick', method to map the data in a high dimensional space before the actual classification step. This pre-step allows this linear model to deal non-linear problems. It is particularly good for classification problems with a large feature space (such as discourse analysis where many features have to be taken into account) because the training time is not long.

*Structured Perceptrons*: a supervised method which returns a structured output using a very basic type of neural network, the perceptron (a linear classifier). It can be useful to allow a better integration of different components of the system because it allows joint learning.

*Recursive Deep Neural Networks (RDNNs)*: also a neural network but deep, that is with many different hidden layers. Deep neural networks (DNNs) have been redicovered by recent technology since before the hardware was not enough powerful to perform this kind of learning. In the last two years DNNs have been increasingly used. The recursive type is a semisupervised method which has been shown to be very good to parse Natural language sentences in Socher et al. (2011).

## **Similarity methods**

*Pointwise Mutual Information (PMI)*: a method to measure the exetnt to which two objects are associated according to their distribution. It is the ratio of the probability of the two elements occuring together to the probability of their independent occurrences. It is used to measure the degree of association between words, thus to extract relevant lexical features for discourse analysis.

## **Mathematical models**

*Integer Linear Progamming (ILP)*: linear programming is a model used to optimize the output of an objective function given some constraints. It can be used to model a wide variety of NLP tasks which require a choice among a set of possibilities forced by some limitations (such as grammatical rules or contextual information). Integer linear programming is a type of linear programming in which at least some of the variables are required to be integers.

### 1.3 Outline of the essay

After having introduced the reader in this section to the theoretical and methodological background necessary for the following discussion, in the next section (§2) we will propose a general overview of the past and state-of-the-art research using discourse relations for various Natural Language Processing tasks.

In the beginning of the chapter (§2.1) we will present the main resources developed by the literature to study discourse relations, that is the RST-DT and the PDTB. We will argue that the creation of these annotated corpora has played an essential role in the development of the field, providing both shared training sets which improved standardisation and comparability across the various approaches. Afterwards (§2.2), we will provide an overview of the main NLP approaches using coherence relations, that is the ones dealing with written text. Firstly (§2.2.1) we will discuss the literature concerned with the basic tasks of discourse relations automatic assignment and its extension to recreate the whole hierarchical structure of the text (discourse parsing). Subsequently (§2.2.2) the main area of applications of the discourse relations detection systems previously discussed will be presented, with some examples from recent approaches using these systems to perform information extraction, attribution detection and summarization. In the last section (§2.3) we will propose a novel (almost) unexplored field of application for discourse relations: speech processing, both from the point of view of recognition and synthesis.

Finally, in chapter 3, first (§3.1) we will summarize the main ideas and results discussed in the previous chapter, then (§3.2) we will propose some final remarks regarding the main challenges that the field of computational approaches to the rhetorical analysis of discourse will have to face in the future with some proposals regarding some directions of the work to be undertaken next.

## Chapter 2

# Computational approaches to discourse analysis using coherence relations

### 2.1 Resources

Possessing the right resources is a factor of fundamental importance for the development of an area of research, especially if the orientation is towards applications. Sharing the same training and testing corpora allows indeed different scholars working in the same field to compare the results of their studies, thus increasing the speed with which the research in the whole area proceeds. Furthermore, shared resources are essential for having a standardization across groups using various theoretical and methodological frameworks. The development of large shared resources annotated with discourse relations in the last 15 years was one of the main reasons for which the area of automatic analysis of the rhetorical structure of discourse was so developed

The first large resource which was developed to study discourse relations was the Rhetorical Structure Theory Discourse Treebank (RST-DT) corpus (Carlson et al., 2001), composed 385 articles from the Penn TreeBank annotated a set of 78 coherence relations formulated according to the RST framework and grouped in 16 macroclasses.

A few years later, however, the importance of this resource for computational discourse was substituted by the much larger Penn Discourse Treebank (PDTB) (Prasad et al., 2007b),(Prasad et al., 2008), a corpus which provides a further level of annotation (the one of discourse relations) on top of the previous ones on the entire the Penn TreeBank.

Although the annotation of the PDTB follows in general a lexically motivated approach to DRs annotation (Webber et al., 2003), this resource was developed with an inclusive criterium. The annotation allows thus test different theories regarding discourse relations (including the RST framework Mann and Thompson (1988)).

More specifically, in the PDTB the sentences in the whole corpus are annotated as either: Explicit (part of an explicit DR), Implicit (part of an implicit DR), AltLex (discourse relations signalled by alternative lexicalizations), EntRel (which identify the cases where only a coherence based on entity was retrieved) or, if no DR is present, as without any relation (NoRel). Each relation takes exactly two argument spans (Arg). As shown in Figure 2.1, the set of relations is organized hierarchically in three levels, with the first one identified by the macrocategories TEMPORAL, COMPARISON, CONTINGENCY and EXPANSION and the second level composed by 16 relations. Each relation in the PDTB is thus annotated with three different levels of DRs.

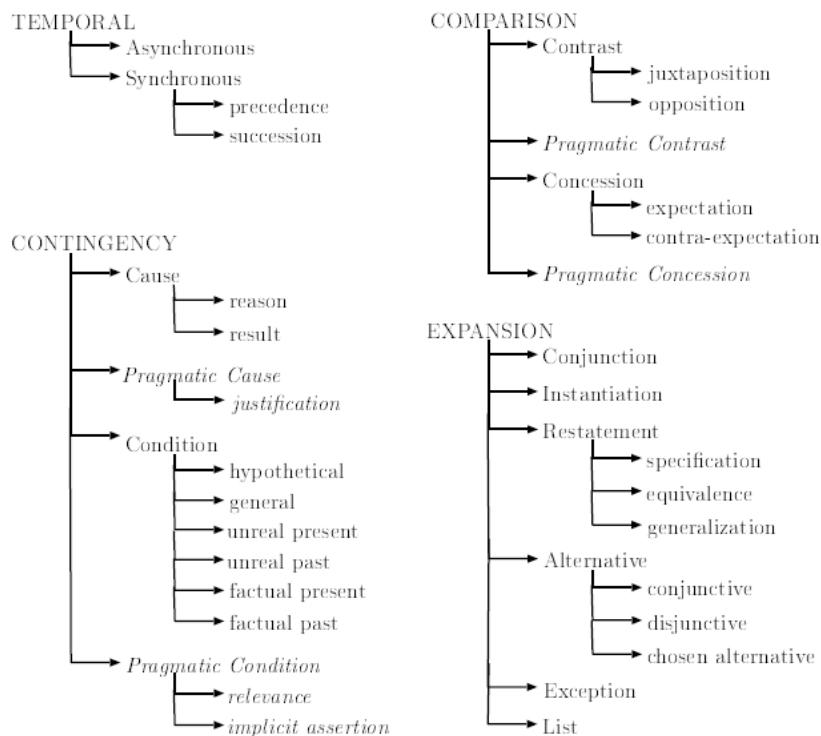


FIGURE 2.1: Hierarchy of sense tags taken from Prasad et al. (2008).

Besides classic DRs, a further annotation level which has been developed for the PDTB is the one of Attribution Prasad et al. (2007a), a relation which cannot properly be defined as a discourse relation, but which can nevertheless greatly affect our interpretation of the text. Attribution, annotated in the PDTB according to the four dimensions sources, types, scopal polarities, and determinacy, is the relation which connects the source of a quotation to the reported speech in the text. We have decided to include this relation

in our discussion because in most of the approaches to automatic discourse relations assignment system this relation is considered along with the other classical DRs. Also, we will see later (§2.2.2) how research on Attribution Relations is currently an interesting and novel field of study which could be very useful for applications.

As a conclusive remark regarding the two resources which constitute the main test corpora for all the computational approaches to discourse relations (which will be discussed in the next section), we must notice how both of them use only Wall Street Journal articles (WSJ). There is thus no variation at all in the type of english they portray, that is the formal business english used in this newspaper. This factor must be taken into account when evaluating the performance of the different systems as well as their portability and scalability across different applications.

## **2.2 Main approaches in the literature: NLU for written texts**

### **2.2.1 Discourse relations assignment and discourse parsing**

In this section we present an overview of the various approaches to rhetorical structure automatic analysis, that is the task of automatically detecting discourse relations in a text along with the possible extension of combining the coherence relations identified to recreate a hierarchical discourse structure of the entire text.

This task, with or without its possible extension, involves a series of steps, each of which is essential for the final result.

Firstly, we must perform discourse tokenization of the text. This means chunking the discourse in discourse units, which will later serve as arguments for the coherence relations. This task is very hard and the literature is still divided on how to solve it. For the sake of the performance some approaches give it for granted, thus using an already tokenized text in order not to affect the performance of the subsequent steps. Discourse tokenization can be performed as a rule-based or classification task to decide if each boundary over words should be regarded as a discourse boundary or not. Some other approaches simply assume plain sentences as the discourse tokens.

Once the discourse has been divided into units we have to identify the discourse connectives (if they are present) and then the arguments of each discourse relation. Generally these substeps are performed in this order, but there are also some approaches which try to further divide these passages in order to have the information coming from both

discourse connective and arguments identifiers to interact before taking the final decision regarding the presence of a DR or not.

Most of the approaches to coherence relations automatic assignments stop here. There are however few studies which also tried to do proper discourse parsing, that is a final step in which the hierarchical discourse structure of the whole text is created by combining the different discourse relations identified.

In the following discussion first we will present the approaches which attempted this task in a supervised fashion. The presentation of these approaches will be describe first the systems which have been tested in the older and smaller RST-DT, and then the ones which have been tested on the larger PDTB. This way of proceeding is motivated by the aim of comparing the results in the literature. At the end of this section we will discuss a few attempts to try to use unsupervised learning for discourse relations automatic assignment.

The first approaches to coherence relations identification were developed relying on the RST-DT. One of the first discourse parsers was developed by Marcu (2000) almost contemporary to the release of the corpus. This parser was based on a decision tree algorithm and was aimed at reconstructing the entire hierarchy of discourse relations in the RDT framework.

After these early approaches, the first system which represented a benchmark in the literature using RST-DT to build discourse analyzers was the one described in Soricut and Marcu (2003). This system, called *Spade*, was aimed at reconstructing only intrasentential coherence relations, thus not the full DRs hierarchy of texts. It addressed both explicit and implicit relations. In particular, in this implementation two probabilistic models were built to reconstruct sentence level parse-trees using both syntactic and lexical features. The system is divided into two steps, discourse segmentation and parsing. Firstly, starting from the output of the Charniak parser, discourse segmentation is performed using syntactic and lexical information to determine the probability that a word could be followed by a discourse boundary. Afterwards the discourse parser takes as input the discourse segmented lexicalized syntactic tree represented with a dominance set (used to infer dominance relationships between trees which could be useful to reconstruct discourse relations). At this point a discourse model is constructed to compute probabilities for different discourse parse trees using a formula which computes Maximum Likelihood Estimation smoothed with simple interpolation. Finally the discourse parser uses dynamic programming algorithm to search the space of all possible parse trees according to the model. This implementation achieves an improvement of about 3% F-score for the segmenter, and about 40 % for the parser compared to the decision based system proposed in Marcu (2000).

The usefulness of lexical and syntactic information for discourse parsers was successively extended in the literature, with further extensions.

In particular, Duverle and Prendinger (2009)'s approach, using a SVM method with a much larger space of features (such as shallow lexical, syntactic and structural (ex. the fact of belonging to the same sentence or to the same paragraph, the measure of the span size and positioning) features), has been shown to outperform Soricut and Marcu (2003)'s implementation achieving, in linear time an F-score of 73.9%. The implementation of Duverle and Prendinger (2009) is also divided into two steps. In this case two SVM classifiers are trained, the first one to decide if there is a discourse connection between two segments, the second one to identify the rhetorical relation which connects the segments previously detected. A bottom-up tree building algorithm is then used to reconstruct the hierarchical discourse structure. The authors justify their choice of using SVMs with the facts that these methods, being maximum margin classifiers, avoid the overfitting at the same time offering more resilience against noisy input (such as discourse). The short computing time is also a major advantage given the big features space they use. This system is then compared to both the early system proposed in Marcu (2000) and *Spade*, showing improvement over both.

Discourse relations extraction systems trained on the RDT-DT show thus a progressive increase over the years achieved mainly by using more features and more flexible algorithms. In all these studies is however reported the still unsolved problem of discourse segmentation which, according to the authors, could constitute a major bottleneck affecting the performance of their systems.

With the release of the PDTB, especially its second version, it was finally possible to have a larger resource to train the automatic DRs assignments systems. The development of such a large resource has also attracted further attention on the problem of rhetorical structure automatic analysis, since the size and comprehensiveness (from the point of view of the theoretical approaches) of this resource made also the task of discourse analysis more challenging.

Also for this reason, most of the approaches which worked on the PDTB focused on trying to solve specific problems (such as only implicit DRs recognition) or only particular components of the system of automatic DRs extraction systems.

Some studies, for example focused only on explicit coherence relations. Pitler and Nenkova (2009), for example, study two types of disambiguation problems connected with this type of relations: the fact that a word can or cannot be a discourse connective, and the fact that a word can mark different types of coherence relations. In particular,



they use MaxEnt with a ten-fold cross-validation to evaluate the usefulness of syntactic features for both. For the first task (discourse connective identification), using only syntactic features they achieve an accuracy 92.25% (with a consistent improvement over the baseline system). By adding interactions between pairs of syntactic feature they are able to reach an accuracy of about 94%. For the second task (identifying which type of relation a discourse connective is signalling) they report a less than 1% improvement (94.15%) by adding syntactic features, remarking however how the same accuracy was achieved by human annotators (which could mean that they are reaching a ceiling boundary). Besides this further confirmation of the usefulness of syntactic features, Pitler and Nenkova (2009) shows also that the task of automatically detecting explicit discourse connectives seems to be tackled correctly by the systems in the current literature, given the high accuracies returned by rather simple implementations such as the one presented in this article.

On the other hand, systems trained on the PDTB whose focus was on recognizing implicit discourse relations did not have the same success.

One of the first pioneering approaches to this task was presented in Marcu and Echihabi (2002), which demonstrated in particular that a word pairs approach seems to be comparatively successful to automatically detect implicit DRs.

On this early work rely both Pitler et al. (2009) and Lin et al. (2009). The second article, in particular, reported quite successful results with a significant 14.1% improvement over the baseline and an overall performance is 40.2% (which set the state-of-art at the time of the publication). To achieve these results the authors trained a MaxEnt classifier aimed at recognizing the second hierarchical level of DRs on the PDTB which uses contextual, syntactic (both from constituency and dependency parsing) and lexical features. As suggested in Marcu and Echihabi (2002), they employ word-pairs information along with contextual and syntactic information on the two potential arguments of the relation.

The state of the art on automatic implicit relations classification was further improved in Ji and Eisenstein (2014) (43.56% accuracy) thanks to a novel compositional distributional semantic approach which, instead of considering only the semantic of the two arguments of a discourse relation, computes also a meaning representation for the entities in the arguments, relying on the assumption that possible coreference relations between the entities could decrease the performance of an implicit DRs classifier. This approach employs recurrent neural networks, in particular in the formalization for parsing natural language sentences proposed by Socher et al. (2011). Thanks to this algorithm the authors are able to recursively combine the different semantics vector representations for both the arguments and the entities (specifically their semantic role) inside the arguments and then combine them as bilinear products. In the evaluation section,

this paper shows how the contribution of entity semantics allows this implementation to reach a higher performance (which supposedly could be even larger in a corpus with more coreference relations).

The approaches to discourse relations assignment trained on the large PDTB corpus presented so far provided only fragmented insights on the task of coherence relations automatic recognition.

Lin et al. (2014) presents the first implementation of an end-to-end discourse parser trained on the PDTB. This system, whose architecture is shown in Figure 2.2, is aimed at recognising both explicit and implicit DRs using a series of MaxEnt classifiers.

The first step in this implementation is aimed at explicit DRs classification and is divided into three further components. Firstly there is a connective classifier (which uses syntactic, contextual and POS features), which identifies connectives from a text. Then an argument labeler is employed to identify the corresponding arguments for a given connective (subdivided in position classifier and then arguments extractor). Finally an explicit connective sense disambiguator provides the sense of a given connective.

In the second step all adjacent sentence pairs within each paragraph are further examined. For each pair of sentences that is not identified in any Explicit relation from the previous step, the parser then classifies this pair into either EntRel, NoRel, or one of the Implicit/AltLex relation types (for which the authors follow the implementation in Lin et al. (2009)).

The third step is dedicated to attribution identification through a final part of the system called Attribution span Labeler (without labelling the 4 dimensions this relation is associated with in the PDTB). In this phase, the parser first chunks the text into clauses, and then for each clause that appears in any discourse relations it checks whether that clause is part of an attribution span.

The evaluation shows that for the explicit connective classifier there is an improvement compared to Pitler and Nenkova (2009). This study shows ths that building an whole end-to-end system for discourse relations classification is possible. From its presentation in Lin et al. (2014) this system has indeed been used by many other research groups.

One recent paper, in particular, has proposed a rather interesting extension to this system. Do et al. (2011) combines indeed the predictions of the discourse parser from Lin et al. (2014) with the ones coming from distributional similarities methods (and in particular PMI), widely employed in computational discourse, to automatically decide if two clauses are in a causal relation or not. Using an integer linear programming

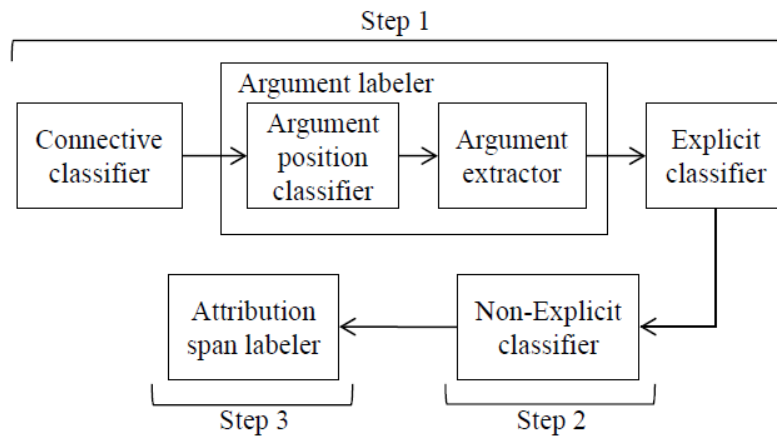


FIGURE 2.2: Pipeline of the end-to-end discourse parser proposed in Lin et al. (2014).

formalization the authors demonstrate how it is possible to successfully "force" the predictions of two independent sources of information to cohere, if they are modelled as linear constraints to an objective function.

Another recent article, following the successes of Lin et al. (2014), has attempted to build an end-to-end discourse parser using the PDTB employing a different methodology. Instead of using MaxEnt classifiers, Li et al. (2014) propose a system able to perform joint learning in an online fashion of all the parameters using a structured perceptron. This system is again composed by two segments: a connective labeller, which identifies discourse connectives and the DR they signal, and an argument identifier. The authors argue that the structured perceptron is specifically used to have a better interaction between these two components (otherwise the model would have had ignored the relations between the connective labeller and the argument labeller) as well as reducing error propagation. The performance of the system on the PDTB is not better than Lin et al. (2014), but is comparable and open to many potential expansions.

After having discussed the main literature regarding discourse relations identifiers learned in a supervised way, we are now going to present the few attempts made so far to use unsupervised learning for the same task.

Marcu and Echiabi (2002) claims a 93% accuracy on recognizing four types of implicit DRs on data automatically annotated using a novel approach. The method they propose to generate their corpus in an unsupervised way consists of using the sentences in explicit discourse relations already annotated and then simply removing the discourse markers from the extracted data. Then Naive Bayes classifiers, which assume that the relation that holds between two spans can be determined on the basis of co-occurrences between words, are trained to distinguish between the different relations.

This approach has however been widely criticized by different studies in the literature, which remarked that without the discourse marker the same discourse relations between the sentences could actually change. In particular, Sporleder and Lascarides (2008) argues that this approach is not valid because it doesn't generalize to new test data and is highly dependent on the type of classifier used. In order to prove this statement, the authors train two classifiers: a Naive Bayes model which uses word co-occurrences to predict a relation; and a BoosTexter-based model which employs a series of linguistic cues. The authors show that while the models were learning from the automatically labelled data, their predictions did not generalise to unmarked examples.

The problem of automatically acquiring new data for discourse relations classification seems thus to remain unsolved so far.

### 2.2.2 Applications

Discourse relations automatic extraction systems, given the fact that they provide a representation of the logical structure of a text, which is essential for its understanding, are used in a wide variety of application.

One of the main uses of coherence relations is naturally in the general field of Information Extraction. (Maslennikov and Chua, 2007), for example, employ DRs to extract the semantic relations between entities as a complement to syntax. Instead of considering only statistical correlation and dependencies between entities as in most of the previous literature, this approach employs also information coming from the recreated hierarchical structure of text, generated using the parser created in the RST framework by Soricut and Marcu (2003).

Another field of application for discourse relations analysis is in Question-Answering systems. A discourse parser would indeed allow a better understanding of the real meaning of complex queries. To this type of applications are connected the ones for Sentiment Analysis and Opinion Mining, where the evaluation of the nuances of the meaning is essential.

In this field, some interesting work is recently being done to create systems which automatically extract quotations and reconnect them to their rightful sources from a text using the Attribution Relations proposed in Prasad et al. (2007a) and now part of the PDTB annotation. An example of this type of research is Pareti et al. (2013), where the authors build an automatic Attribution Relations which identifies reported speech using the k-NN algorithm with 20 features types. This article relies on a more comprehensive definition (Pareti, 2012) to attribution which, starting from the theoretical framework

of Prasad et al. (2007a), is aimed at having a more inclusive approach in order to detect not only direct quotations, but also indirect quotations and opinions. The authors show that supervised approaches outperform the rule based approaches to attribution used in the previous literature.

Finally, another important field of applications for discourse relations is in text generation to evaluate the coherence of automatically generated documents, such as summaries. Already early work (Marcu, 1998) in this field had showed that rhetorical relations can be very useful for summarization.

In the last few years, discourse relations have indeed have been successfully used to assess the quality of summaries. For example Louis and Nenkova (2011) proposed a logistic regression classifiers of general versus specific sentences, under the assumption that general sentences are distinctive of human summaries, but not of the ones generated automatically. In order to give a judgement over the quality of artificial summaries it could thus be useful to have a measure of the generality of the sentences in the text. In order to build such a system the authors use the discourse relations Specification and Instantiation and train their classifier on the PDTB, reaching an overall accuracy of 75%.

Besides the particular implementations presented in this section as examples, we can foresee that with the increase in the performance of discourse relations automatic identifiers there will be a corresponding growth in the use of discourse relations for other applications, given the fact that the systems could then rely on more correct predictions regarding the rhetorical structure of the text.

### 2.3 A new possible direction of research for DRs: Speech processing

- (5) “ Take care of the sense, and the sounds will take care of themselves.”

Lewis Carroll, *Alice’s Adventures in Wonderland*

As suggested by the Dutchess in Example 5, what we really care about whenever we speak is communicating a given message, a meaning. In written text, to make the reader understand which is the correct interpretation, the sense of our message, we use punctuation. Two sequences of words marked by different punctuations could have entirely different senses. The punctuation is thus the key to the right interpretation of the message.

However, when we switch to spoken language, we do not have at our disposition the tool offered by punctuation anymore. The meaning, the correct interpretation of the

sequences of words we produce is entirely encoded by a grammar made of variations in pitch, pauses, intensity, and other features which belong to the area of sound, in one word: prosody.

The field of speech processing is a very challenging and exciting one nowadays in computational linguistics. Besides all the problems faced by NLP applications for written language, speech processing has the further difficulty of the conversion of a discrete object such as words into the continuous dimension of the signal (speech synthesis) and the viceversa (speech recognition).

This means that the most basic task for NLP, that is words recognition, in this field is not given for granted. This task is particularly complicated because spoken language is a much richer and various mean than written text (Shriberg, 2005). Each individual has its own voice and way to speak which affect the pronunciation of words. The context produce noise. Punctuation is absent. Disfluencies fragment the fluency of speech. All these facts constitute major challenges for this field.

In the last years, before facing with the more abstract problem of spoken language understanding, this area of research had thus to deal with simple words recognition and synthesis. However, in recent times, thanks to advances in both software (new machine learning methods such as DNNs) and hardware along with the availability of much bigger corpora, this field has developed rapidly.

Only in very recent times, thus, the problem of actually understanding spoken language has arised in the field. Finally, after many years researchers studying speech strated to care about the sense in as portrayed by the sound. Processing prosody represents naturally a major challenge for this task.

My proposal is that discourse relations could be profitably used to study the "grammar" of prosody. What rhetorical relations are indeed aimed at representing is the sense of the discourse, its logical structure. I believe it could be thus very interesting to study the interrelations of different discourse relations with different types of prosodical structures. Prosodical cues (such as variations in pauses or pitch) are indeed used by speakers to portray the sense of what they are saying.

In order to study this interrelation, we could automatically transcribe a given spoken corpus and then have human annotators to check the correctness of the transcriptions and to annotate the discourse relations in the sentences. Relying on such a corpus, we could then analyze the potential correlations between given discourse relations and their prosodical encoding.

To the best of our knowledge, only one attempt has been done so far to try to connect DRs with speech to study the relation of contrast in the context of a concept-based speech synthesis system (Theune, 2002).

The outcome of a large analysis on the correlations between discourse relations and prosody could be exploited both for speech recognition, for example to improve sentences segmentation and in general spoken language understanding, and for speech synthesis, for example it would be possible to use this "prosodical grammar" to produce more natural synthesized voices.

## Chapter 3

# Conclusions

### 3.1 Summary

In this essay we provided an overview of the uses of discourse relations for the automatic analysis of the rhetorical structure of discourse.

First we proposed a theoretical introduction to the field of computational discourse in general, explaining the challenge of analyzing discourse automatically, and to discourse relations in particular, motivating our choice of focusing on coherence relations with their success in recent research in NLP and with their ability to tackle the main property of coherence, that is that the discourse must have a logical structure, a sense. We then provided a brief introduction to the main methodological tools used to deal with discourse relations automatic assignment.

In the central chapter, we introduced the RST-DT and the PDTB, the two resources which allowed for the research in this field to have such an increase in the last few years. Afterwards we discussed the main research directions in computational analysis of discourse structure using rhetorical relations so far.

In particular, we presented an overview of the various approaches developed for written text, first in the field of discourse relations automatic assignment with and without the extension to try to build a full hierarchical structure for the whole text. In this section we also showed the difficulty of building an end-to-end discourse relations assignment system (for both implicit and explicit relations), with the first example in the literature presented only in 2010. We showed however how more research has been is currently being done in this direction, especially using neural networks or correcting the predictions of the system employing other sources of information.



For this discussion emerged also some of the main issues that discourse relations identifiers have to face in the next future in order to further improve their performance. Firstly we must find a solution for the bottleneck of automatic discourse segmentation, the first essential step in each DRs recognizer which greatly affect the performance of the system. Secondly a need to find new data has been highlighted by a part of the literature aimed at trying to create new data in an unsupervised fashion. Thirdly, much more research is needed in the area of implicit relations classification, on which the performance is still rather low.

We then discussed some of the main fields of applications that discourse relations identifiers could have, such as summarization, question–answering systems, sentiment analysis and information extraction. We provided some examples of the research being done in most recent years, which shows an increasing number of successful approaches using DRs for many different purposes.

We then proposed a new exciting possible direction of research on coherence relations with many potential outcomes: the interaction with the growing field of speech processing. In particular we proposed to study the potential interrelation between discourse relations and prosody (both pitch and intensity and pauses) to infer if some meaning are associated to given prosodic structures. This type of research, given the fact that understanding prosody is one of the main challenges of speech processing nowadays, could greatly improve natural language understanding for recognition. It could also be used in synthesis to produce more natural prosody.

In the next section we are going to discuss some final remarks regarding the challenges that discourse relations automatic extraction have to deal with in the next future as emerged from the previous discussion along with some possible other interesting directions of research which could be exploited in this field.

## 3.2 Final remarks and future work

- (6) “ Alice: ‘Would you tell me, please, which way I ought to go from here?’  
The Cheshire Cat: ‘That depends a good deal on where you want to get to.’  
Alice: ‘I don’t much care where.’  
The Cheshire Cat: ‘Then it doesn’t much matter which way you go.’  
Alice: ‘...So long as I get somewhere.’  
The Cheshire Cat: ‘Oh, you’re sure to do that, if only you walk long enough.’ ”  
Lewis Carroll, *Alice’s Adventures in Wonderland*

The research paths which are currently open for the research on the rhetorical structure of discourse are many. Especially given the lack of a shared theoretical framework, the directions of research on discourse relations could truly be multiple.

However, only in the most recent years, we have seen a flowering of the field, given both by the advances in other areas of artificial intelligence in general and natural language processing in particular, which have allowed this field to tackle more subtle and challenging problems such as automatic discourse analysis. At the same time, the improvements in this field lead to more requests from the industry which is trying to build more sophisticated automatic methods to treat natural languages where a deeper understanding is required.

Besides the connection to the field of speech processing, another area of research in NLP which could greatly benefit from a better analysis of discourse is Machine Translation, another field which has seen a great development in the last years. The successes in this area have indeed brought more attention to the understanding of discourse phenomena which must be analyzed in order to produce a good translation.

On the contrary, an NLP area of research whose improvements could greatly benefit discourse analysis is semantic parsing. One of the major challenges that discourse parsing will have to face in order to truly increase the performance of discourse relations identifiers is having a better understanding of the semantics of the text. Some research groups (Barzilay and Lapata, 2008) are already going in this direction, trying to merge entity-based approaches with discourse analysis. Another challenge would be the one to merge discourse with the world knowledge information, coming from models of Knowledge Representation such as ontologies.

Finally, we need to explore new genres and languages. As previously noticed, the PDTB represents indeed only a very partial perspective of English language, portraying a very formal type of English focused on the topic of business. Some efforts (Tonelli et al., 2010) in this direction have been in the past years, for example to create a corpus of dialogues with the annotation scheme from PDTB.

In some way, as stated by the Cheshire Cat in Example 6 it does not matter exactly where we go, which path we decide to explore first, as long as the study of discourse, the most abstract and subtle level of language, continues. For sure the research on the 'sense', on the logic of what we are saying is going somewhere.

# Bibliography

- Nicholas Asher and Alex Lascarides. *Logics of conversation*. Cambridge University Press, 2003.
- Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16*, pages 1–10. Association for Computational Linguistics, 2001.
- Robert-Alain De Beaugrande and Wolfgang Ulrich Dressler. *Introduction to text linguistics*, volume 1. Longman London, 1981.
- Quang Xuan Do, Yee Seng Chan, and Dan Roth. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics, 2011.
- David A Duverle and Helmut Prendinger. A novel discourse parser based on support vector machine classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 665–673. Association for Computational Linguistics, 2009.
- Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.
- Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, pages 44–50. Association for Computational Linguistics, 1983.
- Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225, 1995.
- Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. *Cohesion in english*. Routledge, 1976.
- Jerry R Hobbs. Resolving pronoun references. *Lingua*, 44(4):311–338, 1978.
- Jerry R Hobbs. Coherence and coreference. *Cognitive science*, 3(1):67–90, 1979.
- Eduard H Hovy. Parsimonious and profligate approaches to the question of discourse structure relations. In *Proceedings of the 5th international workshop on Natural Language Generation*, pages 128–136, 1990.
- Yangfeng Ji and Jacob Eisenstein. One vector is not enough: Entity-augmented distributional semantics for discourse relations. *arXiv preprint arXiv:1411.6699*, 2014.

- Aravind K Joshi and Steve Kuhn. Centered logic: The role of entity centered sentence representation in natural language inferencing. In *IJCAI*, pages 435–439, 1979.
- Min-Yen Kan, Judith L Klavans, and Kathleen R McKeown. Linear segmentation and segment significance. *arXiv preprint cs/9809020*, 1998.
- Sheng Li, Fang Kong, and Guodong Zhou. A joint learning approach to explicit discourse parsing via structured perceptron. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 70–82. Springer, 2014.
- Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 343–351. Association for Computational Linguistics, 2009.
- Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, pages 1–34, 2014.
- Annie Louis and Ani Nenkova. Automatic identification of general and specific sentences by leveraging discourse annotations. In *IJCNLP*, pages 605–613, 2011.
- William C Mann and Sandra A Thompson. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute, 1987.
- William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.
- Daniel Marcu. Improving summarization through rhetorical parsing tuning. In *The 6th Workshop on Very Large Corpora*, pages 206–215, 1998.
- Daniel Marcu. *The theory and practice of discourse parsing and summarization*. MIT press, 2000.
- Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375. Association for Computational Linguistics, 2002.
- Mstislav Maslennikov and Tat-Seng Chua. A multi-resolution framework for information extraction from free text. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 592, 2007.
- Silvia Pareti. A database of attribution relations. In *LREC*, pages 3213–3217, 2012.
- Silvia Pareti, Tim O’Keefe, Ioannis Konstas, James R Curran, and Irena Koprinska. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 2013.
- Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16. Association for Computational Linguistics, 2009.
- Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 683–691. Association for Computational Linguistics, 2009.

- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. Attribution and its annotation in the penn discourse treebank. *Traitement Automatique des Langues, Special Issue on Computational Approaches to Document and Discourse*, 47(2):43–64, 2007a.
- Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. The penn discourse treebank 2.0 annotation manual. 2007b.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. The penn discourse treebank 2.0. In *LREC*. Citeseer, 2008.
- Elizabeth Shriberg. Spontaneous speech: how people really talk and why engineers should care. In *INTERSPEECH*, pages 1781–1784, 2005.
- Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136, 2011.
- Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 149–156. Association for Computational Linguistics, 2003.
- Caroline Sporleder and Alex Lascarides. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(03):369–416, 2008.
- Mariët Theune. Contrast in concept-to-speech generation. *Computer Speech & Language*, 16(3):491–531, 2002.
- Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind K Joshi. Annotation of discourse relations for conversational spoken dialogs. In *LREC*, 2010.
- Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. Anaphora and discourse structure. *Computational linguistics*, 29(4):545–587, 2003.
- Bonnie L Webber. A formal approach to discourse anaphora. Technical report, DTIC Document, 1978.
- Bonnie L Webber. So what can we talk about now? In *Computational Models of Discourse*, pages 331–371. The MIT Press, 1983.