Using Coreferences for Coherence Relations

Holger Schauer
Computational Linguistics Division
Freiburg University
D-79085 Freiburg, Germany
schauer@coling.uni-freiburg.de

Abstract

Although widely acknowledged, work on discourse structure of texts has often not addressed the role of referring expressions. I will argue that any analysis of discourse structure should obey the referential dependencies that can be found in a text. Under this assumption, the resolution of anaphoric expressions can be used as a guideline for automating the derivation of a text's discourse structure. Finally, the proposed algorithm is evaluated.

1 Introduction

It is widely acknowledged that texts are not just a collection of sentences, but have a structure of their own. The relationship between the entities referred to in a text builds up a structure based on *coreference*. In Example (1), the pronominal *"it"* refers to *"the Vaio F190"*, just like the nominal *"the notebook"* or the textual ellipsis *"the batteries"* (of *"the Vaio F190"*).[1] The references of a text establish accessibility constraints for further resolution processes of subsequent anaphoric references, e.g. the *"weight"* in (1-c) should no longer be accessible after processing (1-d). This *referential structure* is reflected in approaches that use e.g. a stack of lists of antecedent candidates as the data structure for reference resolution purposes, like (Hahn and Strube, 1997). However, the referential relationships between objects are usually semantically poor, i.e., finding the antecedent for an anaphoric expression does not add further heavy semantic information to an analysis.

(1) a. $\{$*The Vaio F190*$\}_1$ *comes with convincing equipment.*
b. $\{$*It*$\}_1^{pa}$ *features a DVD-ROM drive and a 6 GB harddisk.*
c. *Still* $\{$*the notebook*$\}_1^{na}$ *has a weight of only 3kg.*
d. $\{$*Its*$\}_1^{pa}$ *14"-display shows very bright colors with clear contrasts.*
e. $\{$*The batteries*$\}_1^{te}$ *last two hours.*

In contrast, work on *coherence* of texts is concerned with semantically rich relations among expressions, typically among clauses and larger text fragments. While the research community does not fully agree on a canonical set of *coherence relations*, these involve significant information like *Cause* or *Evaluation*.[2] For instance, (1-a) might be taken to *Evaluate* (1-b). Usually these coherence relations are taken to stipulate a structural configuration between the *discourse units* they relate: coherence relations are either paratactic, i.e. the related discourse units are on a par, or hypotactic, i.e. one of the units dominates the other (Halliday and Hasan, 1976). E.g., Rhetorical Structure Theory RST (Mann and Thompson, 1988) distinguishes between relations that relate equally important units and such that relate a more important (nuclei) and a less important unit (satellite).

[1] The referee is marked with a lowered index, the type of the referring expression is marked with ^{pa} for pronominal anaphora, ^{na} for nominal anaphora and ^{te} for textual ellipsis.

[2] Coherence relations in this paper are basically taken from taken Rhetorical Structure Theory (Mann and Thompson, 1988) will appear *emphasized* and *Capitalized*.

Such structural prescriptions lead to a tree-like analysis of a text's *discourse structure*.

Recent research on the automated computation of discourse structure has applied so-called *discourse cues*, i.e. words or phrases explicitly hinting at the relation at hand (Marcu, 1998). For example, the conjunction *"still"* in Example (1-c) could be taken to indicate a *Concession* relation between (1-b) and (1-c).

Our on-going empirical work, however, indicates that discourse cues might not be sufficient, cf. Section 4. In addition, while cue-phrases may signal which coherence relation is appropriate, they usually do not determine the target unit to which a new unit should be attached. Therefore, it is necessary to consider further elements of *cohesion*, i.e. linguistic devices that establish a text's coherence (Halliday and Hasan, 1976). While referring expressions are generally seen as one such cohesive device, there is surprisingly little work on their exact role for coherence. This paper will address the question how the structure imposed by coreference relations influences the discourse structure of a text in terms of coherence relations.

The paper proceeds as follows: I first address the notorius difficult notion of "pre-realizational structure" that is associated with Rhetorical Structure Theory and propose a concentration on structural constraints in the text under consideration – focusing on referential constraints. I then describe the typical (discourse) structural configurations that occur and their relationship to referential dependencies. This leads to a proposal of how the resolution of referring expressions can be used to guide the automated derivation of a text's discourse structure. Finally, I evaluate the approach and conclude with a look on related work and further research.

2 Discourse Structure

2.1 Structural Dependencies

The main focus of Rhetorical Structure Theory is the effects that were intended by the author of a text. The structure reflects what was probably of highest concern to him, e.g. a nucleus should contain the information that was more important for the author to convey. RST focuses on a "pre-realizational" structure (Mann et al., 1992, p. 45), i.e. it is not primarily concerned with text phenomena. This decision makes it often difficult to judge the correctness of a structure as the authors intention is seldom apparent, neither to human annotators nor to automated systems.[3] This is especially true for judgments of what might have been of most importance to the author, which is at the heart of the theory.

The result of this focus are analyses that do not reflect *structural constraints in a text*. Focusing on *referential constraints* in this article, the dependency of an anaphoric expression on its antecedent – that it is resolvable and hence interpretable – can hardly be captured in RST.[4] While this is of course not an issue for RST as a theory, it is a prerequisite for any system that wants to account automatically for a text's discourse structure, text understanding systems and generating systems alike. As an example consider the following fragment repeated from Example (1):

(2) a. $\{$*The Vaio F190*$\}_2$ *comes with convincing equipment.*
    b. $\{$*It*$\}_2^{pa}$ *features a DVD-ROM and a 6 GB harddisk.*

In classical RST, example (2-a) could be said to stand in an *Evaluation* relation to (2-b). The definition of *Evaluation* requires that the satellite evaluates the nucleus, see Figure 1[5]. However, this would not capture a structural dependency: the pronominal anapher *"it"* cannot be interpreted correctly without the antecedent, so (2-b) depends on (2-a).

[3] This is already acknowledged in (Mann and Thompson, 1988, p. 246): "During analysis, judgments must be made about the writer or readers. Since such judgments can not be certain, they must be *plausibility judgments*."

[4] Another example are syntactic constraints: as (Webber et al., 1999) point out, the syntactic constraints that are associated with cue-phrases also influence a text's discourse structure.

[5] The depicted structures reflect standard RST schemata. The target of the relations is the nucleus.

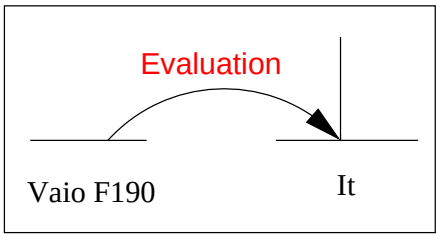

Figure 1: Structure of *Evaluation*

In RST, one could reflect this referential dependency by analyzing (2-b) as giving *Evidence* for (2-a) (see Figure 2), as an *Evidence* relation may be applied when the hypothesis is the most important part, i.e. the nucleus. Such an analysis would neglect the *"convincing"* in (2-a) which is quite an explicit evaluation.

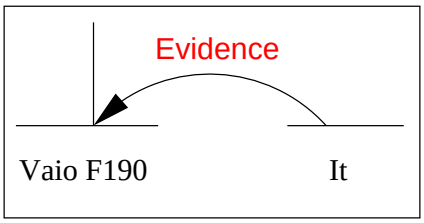

Figure 2: Structure of *Evidence*

We therefore propose to change the focus of theories of coherence relations and to incorporate such structural constraints of a text. E.g., in Example (2) we use an *Evaluation-N*(ucleus) relation that reflects that the nucleus (2) evaluates the satellite (2-b), which captures the structural dependency and allows the analysis with the relation which we consider most appropriate.

2.2 Coreferences and Discourse Structure

Using the outlined notion of structural dependencies as a basic guideline, we analyzed a set of eleven German texts (cf. Section 4), in order to determine the structural configurations involving coreference relations. The coherence relation set we used was basically classical RST, with modifications to incorporate the divergences mentioned in the previous section.[6]

[6]We also follow RST in considering clauses as the elementary unit size, i.e. when a sentence consists of a main clause and a subordinate clause, both are treated as elementary units. By recursively grouping elementary units, complex discourse units spanning several sentences may result.

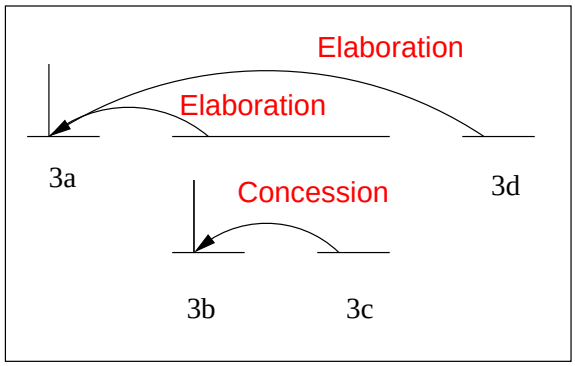

Figure 3: RST analyses for Example (3)

Basically, we found that when two elementary discourse units contain co-referring expressions, they are usually connected by a coherence relation which subordinates the anaphoric unit. In the simplest case, the relation is one of *Elaboration*, in case of textual ellipsis usually due to an *Object-Attribute* relation between the entities. However, further linguistic cues or inferences might give rise to semantically "richer" relations.[7] As I am mostly concerned with the structural configurations, accounting for these further inferences will not be an issue here.

When more than one coreference relation is involved, issues become more complex. For instance, consider the following example:

(3) a. $\{$*The LC90S*$\}_3$ *by Panasonic is a 19"-display.*
b. *Although* $\{$*its*$\}_3^{pa}$ *screen size of 482 mm corresponds to a conventional 21"-monitor,*
c. *considerably less space is required.*
d. $\{$*The device*$\}_3^{na}$ *can be attached to a video-card via a USB-connector.*

Obviously, the nominal anaphor *"the device"* in (3-d) requires an antecedent. One possibility would be the pronominal *"it"* in (3-b), leading to a resolution to *"the LC90S"*.

[7]Section 4 shows that these cues or inferences are sometimes strong enough to override the basic structural configuration, resulting in a multi-nuclear configuration or even a subordination of the antecedent bearing unit.

However, this is not reflected in the discourse structure that seems most appropriate, cf. Figure 3. The topic of (3-b) and (3-c) (the size) is not further elaborated on in (3-d), so one might say there is a mini-segment boundary between these two sentences. Hence, it would also be not correct to analyze (3-d) as an *Elaboration* of (3a-c), because (3-d) elaborates only (3-a).

(4) a. *Apple presented a new PDA,* $\{$*the MessagePad 130*$\}_4$.
b. $\{$*The new display*$\}^{te}_{4,5}$ *has a surface that is more durable than in prior versions.*
c. $\{$*The background lighting*$\}^{te}_5$ *can be controlled with a switch.*

Example (4) illustrates a preference for local attachment: in (4-c) *"the background lighting"* is a textual ellipsis to the *"the new display"* which is in turn a textual ellipsis to the *"Message Pad 130"*. *"The new display"*, however, introduces a new discourse entity that is distinct from – though part of – the *"Message Pad"*. So, this time unit (4-c) is subordinated to (4-b) (cf. Figure 4), because the referential dependency of *"the background lighting"* should be resolved as local as possible. In contrast, in Example (3) there was no such mediated dependency between the anaphoric expressions in (3-d) and (3-b).

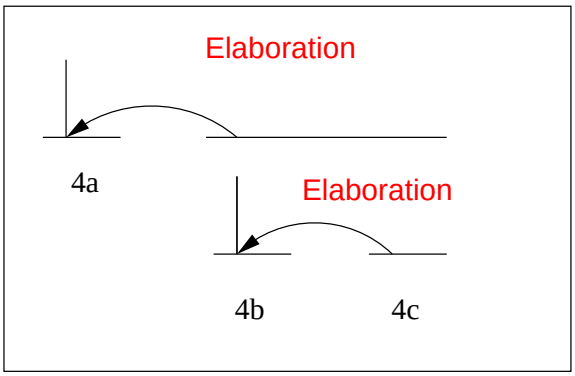

Figure 4: RST analyses for Example (4)

Summing up, the structural configurations that we typically found illustrate a preference to connect a new unit to the highest node (Example (3)) that provides direct antecedents for the referring expression in the new unit (Example (4)).

3 From Coreferences to Coherence

The configurations described in Section 2.2 naturally lead to a combined account of deriving a text's discourse structure and resolving its referring expressions. Basically, the algorithm uses the successful resolution of anaphoric expressions as a guideline for determining the target unit to which a new unit should be connected which in turn restricts the set of units which should be searched for resolving further referring expressions.

The algorithm (given in pseudo code in Figure 5) assumes the availability of several capabilities for anaphora resolution. First, for a given unit candidate (a clause) a set of noun phrases needs to be identified that may be anaphoric expressions (`anaphoric_expressions`), basically pronouns and definite nominal phrases. Second, some resolution process (`match(ana_cand,ante_list)`) is necessary that checks whether an anaphoric expression can be resolved in a given list of possible antecedents. Such a process may make heavy use of an underlying domain knowledge-base (Hahn et al., 1996). Third, while not essential to the algorithm, we assume an ordering function on the antecedent lists in line with Centering Theory (Grosz et al., 1995) (`centers_forward(clause)`).

The basic data structure to operate on is a `node` of a tree. A node consists of one data field and a list of child nodes. Leaf nodes, which have empty child node entries, represent elementary discourse units. The data field holds a list of nominal expressions that are available for coreference resolution at that node.

The algorithm now loops through all clauses of a text, building up both the tree and the antecedent lists incrementally. Basically, whenever a new clause has to be considered, the right frontier of the tree is checked for plausible antecedents: first the lowest rightmost node is checked (`lowest_right_node(tree)`) for possible antecedents, if there is none its predecessor on that rightmost branch of the tree is checked

```
tree := tree(centers_forward(first(clauses)), NIL)
clauses := rest(clauses)
forall clause := clauses do
  ana_nodes := array of lists of nodes.
  forall ana_cand := anaphoric_expressions(clause) do
    node := lowest_right_node(tree)
    while node do
      if match(ana_cand, ante_list(node)) then
         ana_nodes[ana_cand] := append(ana_nodes[ana_cand], node)
      fi
      node := predecessor(node)
    done
  done
  target_node := find_highest_node_matching_all(ana_nodes)
  /* found at least one antecedent node */
  if target_node then
    /* connect new unit to old node */
    connect(target_node, tree(centers_forward(clause), NIL))
  fi
done
```

Figure 5: Algorithm integrating Coreferences and Discourse Structure

(`predecessor(node)`). Now, in order to determine the target node to which the new unit should be connected to, one has to find for all anaphoric expressions (the `ana_cand`) all nodes on the right frontier that contain an antecedent to the anaphoric expressions in the new unit.

When all antecedent nodes are determined, the highest node that provides antecedents for all resolvable anaphoric expressions in the new unit is taken as the target node (`find_highest_node_matching_all`), in accordance with the discussion in Section 2.2.

If a new unit contains no referential expression then the algorithm makes no prediction. If the node to connect to has been found, the new unit is connected to it, i.e. the new unit is established as a satellite to the target unit. This means that the new unit opens a new right-most branch and hence becomes the lowest-right node of the tree. So, the new right frontier consists of the newly attached unit (with its antecedent list), the modified node (still bearing the same antecedent list) and its predecessors.

4 Evaluation

We tested the presented algorithm on the already mentioned set of eleven German texts with roughly 2800 tokens. These texts were taken from two different domains: eight texts were reports from the information technology domain and three reports were taken from the German newspaper "'Die Tageszeitung"'. The texts were annotated manually using RSTTool (Marcu et al., 1999) that we extended to allow annotation of cue-phrases and co-referring expressions. Besides the divergences introduced in Section 2.1, we tried to stick as close as possible to the original RST definitions.[8]

The texts were segmented in 294 elementary units, leading to 272 relations and 220 additional complex units, resulting in 514 units as a whole. In a first step, we tried to determine the amount of cue-phrases, us-

[8]The analyses were performed in joint work with a student that I instructed. As a result, I currently have no data on the inter-rater reliability (kappa-statistics (Carletta, 1996; Marcu et al., 1999)), due to the lack of time to train more annotators and/or to analyze the texts independently of one another.

ing a very simple approach: basically, whenever we found it reasonable that our choice of some relation was probably due to a phrase, we counted it as a cue-phrase. For the 272 relations, we identified a total of 71 cue-phrases. However, some relations such as *Joint* make very weak semantic contributions and are probably never explicitly marked by cue-phrases. In our case, 31 such relations could be subtracted. This still leaves 272-31-71=170 (62.5%) of semantically rich coherence relations that are unmarked ! This shows that at least for some texts further indicators of coherence need to be taken into consideration.

We then tried to determine the role of coreferences. We identified a total of 234 co-referring expressions, consisting of 61 pronominal anaphora, 52 textual ellipsis and 121 nominal anaphora. We also identified 30 relative pronouns, which we also considered for evaluation, arriving at 264 expressions examined.

We then manually evaluated the outlined algorithm, i.e. tried to measure in how many cases the algorithm predicts a structure that matches the structure we felt most appropriate and in how many cases the prediction was incorrect.

Obviously, when a new unit contains no co-referring expression, the algorithm makes no prediction. This is the case for 88 units out of 294 new units (29.9%). 28 of those units contained cue-phrases, while the remaining cases seem to require rather complex inferences to compute the target unit and their relation.

The algorithm determined a node to connect a new unit to in 206 cases (out of 294 elementary units). This number is smaller than the number of co-referring expressions because the new unit that needs to be connected can contain more than one co-referring expression. In 125 cases (60.7%) the prediction was correct, i.e. the predicted target node was the one we also found most appropriate during our annotations.

In 81 cases (39.3%), however, the prediction was incorrect. Of these cases, 52 were due to cue-phrases. Interestingly, 48 out of these 52 are intra-sentential cases, i.e. the related units are clauses within a single sentence, cf. Example (3-b) and (3-c). That such intra-sentential phenomena need to be accounted for prior to coreferences seems to be an intuitive result. When one first accounts for this intra-sentential discourse structure and only then considers the coreference relations, the prediction of the target unit, to which this complex unit should be related to, is correct in 42 out of these 48 cases. Thereby the algorithm would correctly predict 167 cases or 81.1%.

This result might be interpreted in two ways: it could be taken as a hint that the minimal size of discourse units should be entire sentences. Or it might indicate that intra-sentential coherence relations should not be seen as restricting the availability of expressions for further anaphora resolution.

5 Related Work

We mainly applied Rhetorical Structure Theory (Mann and Thompson, 1988) as the basic vehicle of our analyses. We set out to overcome its lack of capturing referential dependencies in texts, in order to arrive at an automatic derivation of a text's discourse structure based on the resolution of co-referring expressions.

(Vonk et al., 1992) argue that overly specific definite noun phrases may signal segment boundaries. Their data also suggests that there is a tendency to use pronouns when other means to signal thematic shifts are available. The structural configurations that we found and the evaluation of the derived algorithm indicate that such thematic shifts occur quite frequently without overly specific noun phrases.

(Corston-Oliver, 1998) enlarges the cue-phrase approach of (Marcu, 1998) in a vein similar to ours. However, how several coreference relations interact with the resulting discourse structure, is not spelled out. Also, in his approach, it remains unclear how the correct target node to attach to is identified.

Segmented Discourse Representation Theory (SDRT), as described in (Asher, 1993),

provides a framework of discourse structure which interacts with referential accessibility constraints. Asher does not rely on coreferences for establishing target units; instead the derivation of a coherence relation (and thereby of the target unit to connect a new unit to) relies on rather abstract connections between "events". While recognizing coreference relations certainly also requires domain knowledge and inference capabilities, recognizing connections between events seems an even more impossible task. The presented approach is hence more light-weight in nature.

(Webber, 1991) argued for resolving discourse deictic references along the right frontier of a discourse structure. For establishing the structure two basic operations on trees are employed. But (Webber, 1991) makes no clear commitment when to choose which operation and how the resolution of references influences the tree's construction. (Webber et al., 1999) apply the same basic two operations for incorporating cue-phrases in a grammar-driven approach to discourse structure. Our algorithm could be integrated with their approach, to account for the cases of intra-sentential phenomena discussed in Section 4. My approach is, however, incompatible with Webbers approach in one crucial point: the operations from (Webber, 1991) can not account for the structures discussed in Section 2.2. After adjoining (3b-c) to the root in Example (3), one can attach or adjoin (3-d) only to (3b-c) or to (3a-c). This amounts, however, to a continuation of either segment, against which we have argued: (3-d) elaborates only (3-a).

Finally, recent work on coreference and discourse structure has shown that the resolution of referential expressions can considerably benefit from following a text's discourse structure (Cristea et al., 1998; Cristea et al., 1999), basing their analysis on original RST. As the proposed modifications to RST (see Section 2.1) exactly aim at referential dependencies, I would expect to find even larger benefits for anaphora resolution with analyses that were made according to the proposed guideline.

6 Conclusion

We argued against the pre-realizational view of Rhetorical Structure Theory that brings with it numerous problems. One is it's lack of reflecting structural dependencies of a text. As the resolution of referential dependency is a prerequisite to account for by any automated system, we proposed to focus the analyses of a text's discourse structure on structural dependencies.

We then presented an algorithm that is able to determine the discourse structure of a text by using successful resolution of referring expressions as a guideline for picking the target units to which new units should be connected to. The evaluation showed correct predictions of a text's discourse structure in 61%. The algorithm should be extended with means to handle cue-phrases (e.g. (Webber et al., 1999)), which is straight-forward: basically the proposed algorithm has to be delayed until the (intra-sentential) cue-phrases have been accounted for. Such an extension would result in nearly 81% correct predictions. The combined algorithm will be implemented in the text understanding system SynDiKATe which already provides means for handling referential relations of object-identity and textual ellipsis (Hahn et al., 1996).

As the aim of the work is to augment the system to recognize a text's discourse structure, finding the correct node to attach to is only part of the work. For computing the correct relation further linguistic cues and/or inferences need to be determined and incorporated.

Acknowledgments

The author is a member of the Graduate Program on *Human and Machine Intelligence* at Freiburg University, which is funded by *DFG*. I would like to thank the staff at the Computational Linguistics Lab, Freiburg, especially the students that helped me gathering the data and also the three anonymous reviewers for their comments.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.

Jean Carletta. 1996. Assessubg agreement on classification tasks: The kappa statistics. *Computational Linguistics*, 22(2):249–254, June. Squibs and Discussions.

Simon H. Corston-Oliver. 1998. Identifying the linguistic correlates of rhetorical relations. In M. Stede, L. Wanner, and E. Hovy, editors, *Proceedings of the COLING-ACL'98 Workshop 'Discourse Relations and Discourse Markers'*, pages 8–14. Montreal, Quebec, Canada, 15 August 1998. New Brunswick, NJ: Association for Computational Linguistics (ACL).

Dan Cristea, Nancy Ide, and Laurent Romary. 1998. Veins theory: A model of global discourse cohesion and coherence. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics*, pages 281–285, Montreal, Canada, August.

Dan Cristea, Daniel Marcu, Nancy Ide, and Valentin Tablan. 1999. Discourse structure and co-reference: An empirical study. In *The ACL99 Workshop on Standards and Tools for Discourse Tagging*, pages 48–57, June.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Udo Hahn and Michael Strube. 1997. Centering in-the-large: computing referential discourse segments. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics.*, pages 104–111. Madrid, Spain, July 7-12, 1997. San Francisco, CA: Morgan Kaufmann.

Udo Hahn, Katja Markert, and Michael Strube. 1996. A conceptual reasoning approach to textual ellipsis. In W. Wahlster, editor, *ECAI'96 – Proceedings of the 12th European Conference on Artifical Intelligence*, pages 572–576. Budapest, Hungary, August 11-16, 1996. Chichester: John Wiley.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*, volume 9 of *English Language Series*. London: Longman.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: toward a functional theory of text organization. *Text*, 8(3):243–281.

William C. Mann, Christian M. Matthiessen, and Sandra A. Thompson. 1992. Rhetorical structure theory and text analysis. In William C. Mann and Sandra A. Thompson, editors, *Discourse Description*, Pragmatics and Beyond, pages 39–78. Benjamins Publishing Company, Amsterdam/Philadelphia.

Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. 1999. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL'99 Workshop 'Standards and Tools for Discourse Tagging'*, pages 48–57, University of Maryland, USA, June 1999. New Brunswick, NJ: Association for Computational Linguistics.

Daniel Marcu. 1998. A surface-based approach to identifying discourse markers and elementary textual units in unrestricted texts. In M. Stede, L. Wanner, and E. Hovy, editors, *Proceedings of the COLING-ACL'98 Workshop 'Discourse Relations and Discourse Markers'*, pages 1–7. Montreal, Quebec, Canada, 15 August 1998. New Brunswick, NJ: Association for Computational Linguistics (ACL).

Wietske Vonk, Lettica G.M.M. Hustinx, and Wim H.G. Simons. 1992. The use of referential expressions in structuring discourse. *Language and Cognitive Processes*, 3/4(7):301–333.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse relations: A structural and presuppositional account using lexicalised TAG. In *Proceedings of the Meeting of the Association for Computational Linguistics*, University of Maryland, College Park, MD, June 1999. Association for Computational Linguistics.

Bonnie Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Natural Language and Cognitive Processes 6(2), January 1991, pp. 107-135.*, 6(2):107–135, January.