Discourse Analyses Correction using Anaphoric Cues

Violeta SERETAN
Language Technology Laboratory, University of Geneva
2, rue de Candolle
CH-1211 Geneva 4, Switzerland
Violeta.Seretan@lettres.unige.ch

Abstract

This paper presents a method for correcting arbitrary (possibly wrong or inadequate) discourse structure analyses, which is based on the constraints over the discourse structure configuration imposed by the anaphoric relations in text. The approach taken underlies on concepts and ideas from Veins Theory (VT; Cristea, Ide, Romary, 1998) as a theory of global discourse cohesion. The method, currently under implementation, will be evaluated against a corpus of texts previously corrected manually.

1 Introduction

The discourse structure analysis is a task of major importance in the more general context of discourse understanding and is the basic prerequisite in the processes like text summarisation and text alignment. Moreover, discourse structure information can significantly contribute to improving the efficiency of, for example, anaphora resolution, information retrieval, and information extraction systems. By decoding the hierarchical organisation of pieces of text, the relations between them, and their relative importance, it can be better accounted for: which are the most important parts of a text (Marcu, 1999), which is the part of the text that corresponds to a translated part of text, in which parts of text the antecedent of an anaphor should be looked for (Fox, 1987; Cristea et al., 2000), where in text it is more likely to find a information of a given type, etc.

Achieving an accurate interpretation for a text is a difficult matter that relies, firstly, on the assertion that the text to analyse is coherent; that is, its author succeeded in producing an intelligible, logically integrated text. Further, it depends on the adopted theoretical discourse structure model, which can offer varied granularity of discourse segmentation and multiple discourse relation classifications, these factors influencing crucially the analysis process.

It is unanimously agreed that the abstract structure of a text is like that of a tree, in which the leaves represent elementary discourse constituents and the nodes represent the relations between parts of text. Provided that, the discourse theories should minutely define, in order the analysis process to be unambiguous, at least: what is the elementary discourse unit, when does a relation hold between the spans of text, and how the constituents of a relation are delimited in the text.

It is not often the case that, first, a given discourse is coherent. An interpretation process should then, at least, provide valid partial analyses for the pieces that have an internal coherent structure, and should deal with cases when the text is, by its nature, ambiguous. Nor it is the case that a human is able to decode a text structure if the above-mentioned definitions are incomplete or imprecise.

Human-generated analyses exhibit a higher degree of uncertainty in what concerns the identification of the relations and the identification of (the size of) relations' constituents, than in what concerns the size of elementary units, or the relative importance of the constituents (Marcu et al., 1999).

Automatic analysis methods are mainly based on superficial information, like text formatting and punctuation, and the use of markers or cue-phrases indicating rhetorical relations (as for example the markers "*First*", "*Moreover*", "*Besides*", "*That's why*" etc.) (Sanders et al.,

1992; Knott and Dale, 1994; Marcu, 2000). They are obviously limited, but their (partial) results can significantly contribute to a final derivation.

We present a semi-automatic method for the correction of arbitrary - manually or automatically produced - derivations. We propose the types of structural modifications to be applied on the partial or distorted output so that the proposed structure exhibits a higher degree of reliability, in terms of discourse coherence. That is, the new structure configuration satisfies the constraints that are imposed by the use of references in text: it is such that a specific hierarchical relation *exists* between two spans *whenever* these ones are related by a reference from one to another.

During the correction process, the user intervention is required (when, for example, subtrees are extracted from a node of a structure tree and adjoined to another) and an answer to a question like "does the relation *relation_name* hold between the units *unit1* and *unit2*?" or "Does the following units have equally important roles, relatively?" is needed in order to allow a modification to be performed.

The experimental results obtained by applying the correction method on a collection of initially analysed texts will be evaluated by comparison with the manual corrections of the same analyses.

Section 2 of this paper presents the main concepts and ideas of Veins Theory, which represents the fundamentals of our approach that will be further described in detail in Section 3. Section 4 contains the description of the correction method and Section 5 presents several details about its implementation, functioning and evaluation. Finally, we outline the conclusions and point out the related work.

2 Veins Theory

We based our approach on the recent Veins Theory (VT; Cristea, Ide, Romary, 1998) which offers a model of global discourse cohesion that encompasses ideas from several discourse theories, the most influencing and currently adopted: Fox (1987), Centering Theory (CT; Grosz, Joshi, and Weinstein, 1995), Rhetorical Structure Theory (RST; Mann and Thompson, 1988).

Evolving from Fox's idea that references are strongly related to discourse structure (that is, cohesion of discourse - or the use of referring expression and other discourse connectives – is explained in accordance with the discourse segments from the attentional focus) and generalising the applicability of CT rules from local (adjacent units) to global discourse, VT uses a simplified, RST-based model of discourse and the RST distinctions between nuclear and satellite nodes (i.e. the most vs. the least important constituent of a rhetorical relation) in order to identify the main threads of discourse, its "veins". The veins of discourse are sequences of elementary discourse units arbitrary distant in the text but close following a hierarchical path (which depends on nodes' nuclearity) and that form an abstract of the text at different levels of detail that is internally coherent and self-sufficient in the understanding independently from the whole text.

The main assumption of VT is, consequently, that the references from a given unit are preferably to units on the same vein, each unit having so associated a domain of referential accessibility over the discourse structure tree.

2.1 The "vein" definition

The definition of vein follows a top-down manner, from the root to the terminal nodes, and uses the notion of *head* which is defined for each node, as follows:

- the *head* expression for a terminal node is the node's unit label itself;
- the *head* expression for an internal node is the concatenation of head expression of its nuclear daughters.

If a node has the vein expression *v*, then the computation of vein expression for each daughter node of it is done as following:

- if the node is a nuclear left daughter, its vein expression is the same as the parent's vein expression (i.e., *v*);
- if the node is a nuclear right daughter and it has a non-nuclear left-sibling, then the units labels from its sister's head (marked in some way) are added to the unit's vein expression inherited from the parent, in an order that follows the linear order of their appearance in the text;

- if the node is a left satellite, its head expression is added to the vein expression inherited from its parent, respecting the initial units ordering;
- if the node is a right satellite, than it inherits the parent's vein *without* the units that are marked, and its own head expression is added to it, respecting the initial units ordering.

Note that the VT discourse model is that of an binary tree (equivalent to a RST tree), and the vein definition makes abstraction of the relations' names.

The intuition behind the vein definition is that a unit is related to these preceding units that are the most important in text (the nuclei, hence heads); that, in the case the discourse in not left-polarised, the preceding satellite is also needed in the understanding of the right nucleus, and, finally, that the accessibility into the units of this satellite is further blocked from units that are in turn right satellites descendents of that nucleus (the interposition of a nuclear node blocks the further reference between satellites).

2.2 Vein resolution

With respect to the resolution of anaphora on their units' vein (situation that we henceforth call *vein resolution*), VT distinguishes between two cases: *direct* and *indirect references*.

- ***Direct reference***

The unit of the closest antecedent of the anaphor is situated on the same vein as the unit of the anaphor (the unit of the closest antecedent is contained in the vein of anaphor's unit).

- ***Indirect reference***

The unit of the closest antecedent of the anaphor is not situated on the vein of the unit of the anaphor, but *a further* antecedent from the same coreference chain *is* situated on that vein.

VT considers that coreference relation induces equivalence classes over the set of discourse entities, that's why it considers indirect references as a case of vein resolution.

Indeed, the understanding of the anaphor is still possible following a way back on its coreferential chain, and the distance on the vein is explained by the use of the "proximity convention" in coreference annotation (that is, the reference is marked back between an anaphor to that coreferring element appearing the most recently in discourse, despite the fact that another such element may precede it that is structurally more strongly related to the anaphor).

The references that are not resolved on the vein are supposed to be of pragmatic nature, and understandable outside the discourse. Profs on empirical evidence on VT claims can be found in (Cristea et al., 2000) and (Ide and Cristea, 2000).

The next section is based on the idea that arbitrary analyses may contain subtrees whose configuration do not reflect a hierarchical relation between two units that *are* actually connected by a reference from one to another. A chain of coreferences clearly states the existence of a hierarchical relation between the units involved, and indicates a vein of discourse. Conversely, the exceptions of vein resolution *could* indicate such misconfigured subtrees and may provide help in structure recovery.

3 The approach

Following the definition of vein and the VT main assumption of vein resolution of anaphora, it is expected that in a correctly built discourse structure, all anaphora (that are not of pragmatic nature) will be resolved on the corresponding veins. This means that the obtained discourse interpretation is in agreement with the prescriptions on text coherence that pertain to the text cohesion, under VT. The units of text that assure the cohesion of text, that is, those containing the specific referential chains, correspond indeed to the units structurally connected by the corresponding veins. Otherwise, if a reference goes outside a vein, it can be the case that a unit misses from the vein even if the unit should, by cohesion, belong to it.

The question then arises whether the current structure configuration is coherent: whether it corresponds or not to a valid interpretation of the discourse.

Each time, then, that an exception occurs in the vein resolution of references in the structure, we verify if the structure in cause was correctly build, and in the case we find ambiguities (pertaining to nuclear roles) or invalid associations of constituents spans, we propose

minimal structure modifications aimed at recovering the vein connection between the units involved.
The structure obtained may not be completely corrected, in the sense that it is still uncertain whether it corresponds or not to the interpretation the author intends for the text, but at least it accommodates the structural restrictions imposed by the references.
This section continues with the presentation of the ambiguity factors that during analysis may lead to inappropriate or invalid discourse interpretations. We provide an example from the corpus we studied in which such factors contribute to producing an invalid analysis, whose misconfiguration was signalled by a vein resolution exception.

3.1 Factors leading to wrong analyses

As pointed out in the Introduction, ambiguity and uncertainty in discourse analysis arises mainly from the incomplete and imprecise indications on the determination of *edus* (elementary discourse units) and relations between textual spans: when does a relation hold between two textual spans and how its constituents are delimited in text. Considering the discourse segmentation in *edus* a secondary issue in analysis[1], we consider that in a particular step of analysis two types of choices crucially intervene in the quality of the result: the choice of relation's constituents, and the choice of relation's type.
The first type of choice induces a structural ambiguity: one can associate the spans a, ß and the obtained structure with the span γ: ((a, ß), γ), or the spans ß and γ and then the span a with the obtained structure: (a, (ß, γ)), so that after several steps of analysis two spans that should have been linked together by a relation will be found in the resulting structure arbitrarily distant one from another. A relation's constituent is incorrectly identified as a sub-part or a super-part of itself.

[1] The segmentation does play a role on the structure "correctness" with regard to the vein resolution, but only when considering a too fine segmentation granularity. Passing to a higher granularity will conserve the vein resolution.

The second type of choice influences the nuclearity of the nodes: one can choose either an inappropriate relation, that does not actually holds between two spans, or one from a possible set of relations that could hold which is not the most adequate. Moreover, the initial RST assignation with nuclear roles is uncertain for several RST relations; it is not clear whether the first or the second argument is a nucleus, e.g. for *consequence* and *purpose* relations (Marcu, 1997).

3.2 Example of invalid analysis

We provide an example from the corpus we studied in which we found that uncertainty factors have led to an invalid analysis. The misconfiguration is detected when verifying the vein resolution of anaphora in text. Compare the figures 1 and 2, which represent two variants of analysis for the following text (in which is also shown the initial segmentation in units labelled from 1 to 8):

(1) BURNS FRY Ltd. (Toronto) --
(2) Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm.
(3) Mr. Wright resigned as president of ***Merrill Lynch Canada Inc****., a unit of Merrill Lynch & Co.,*
(4) to succeed Mark Kassirer, 48,
(5) who left Burns Fry last month.
(6) A ***Merrill Lynch*** *spokeswoman said*
(7) it hasn't named a successor to Mr. Wright,
(8) who is expected to begin his new position by the end of the month.

Figure 1 represents the initial structure analysis, as it was proposed in the manually annotated corpus (the nuclear nodes are marked with an underlying line). The reference to "Merrill Lynch" from unit 6 to unit 3 is an exception from the point of view of vein resolution: its antecedent, "Merrill Lynch Canada Inc.", is found in unit 3 that doesn't belong to the vein expression of unit 6 ("1 2 6 7").
Actually, the annotator has chosen to group spans 2 and 3-5 first and then to adjoin span 6-8, while, as discussed in the section 3.1, one alternative could have been to group first spans 3-5 and 6-8 and then to group it with unit 2. There was an ambiguity as to whether the

daughter of span 6-8 is the span 2-5 or its sub-part, 3-5: the annotator can associate ((2,3-5),6-8), vs. (2,(3-5,6-8)).

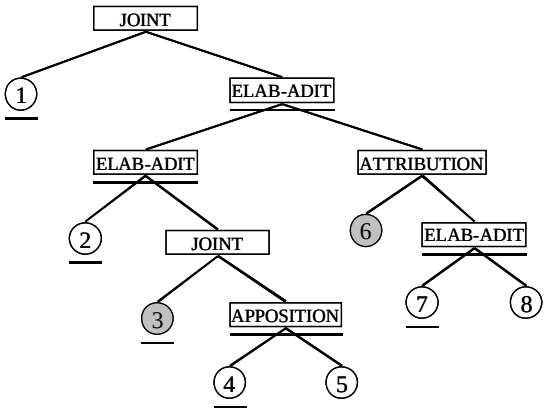

Figure 1 Initial structure configuration for the text.

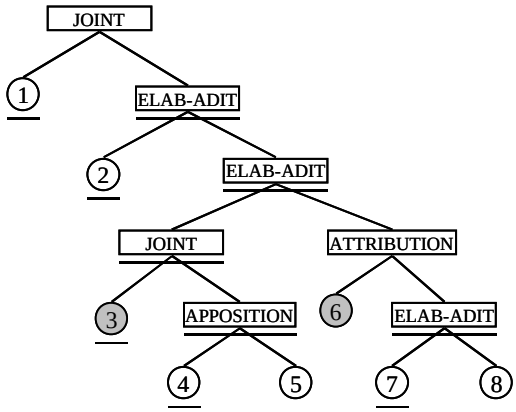

Figure 2 The proposed structure configuration for the same text.

Figure 2 shows the structure obtained by following the last alternative. In the new structure configuration, the exception from unit 6 to unit 3 is a direct vein reference: the vein of unit 6 ("1 2 3 4 6 7") includes unit 3.

Actually, the last alternative is the right one, since the span 6-8 *elaborates* (is in relation of *elaboration-additional* with) span 3-5 only and not the span 2-5: its topic ("it hasn't named a successor to Mr. Wright ") elaborates the topic of 3-5 ("Mr. Wright resigned as president of Merrill Lynch Canada Inc…."), not the topic of span 2-5 ("Donald Wright, 46 years old, was named executive vice president…"). Choosing the last alternative is consistent with the satisfaction of the Compositionality Criterion (Marcu, 1997)[2]: this relation also holds between the salient units 3 and 7, while, in the initial structure, this criterion is not obeyed: the relation doesn't hold between units 2 and 7.

The new structure configuration provides a valid interpretation for the span 2-8, whose misconfiguration in the old structure was signalled by the exception from unit 6 to unit 3.

In what follows we will describe a method aimed at repairing an initial discourse analysis, which uses hints of referential nature and provides the necessary correcting operations on the tree-like structure.

4 The correction method

Given an arbitrary structure analysis, qualified as unreliable or questionable, several modifications are made in the places suggested by the exceptions in the vein resolution, such that the new structure configuration reflects a better interpretation of the entailed (pieces of) text.

The procedure takes into account the types of choices that are possible in a single step of structure derivation and proposes a series of modifications that, if validated by the user, contributes to the correction of the structure affected by the given step of analysis.

We call this process local correction, since only a substructure of the initial structure is recovered, namely that whose possible misconfiguration is indicated by the existence of an exception.

Successive local corrections are applied to the initial structure, linearly and in a bottom-up manner, whose partial results integrate and contribute to repairing the whole structure.

The following sections describe this process in more detail.

4.1 Basic correcting operations

The basic correcting operations on the structure tree concern nuclear role modifications (presuming that from the set of relations that may hold, not the most appropriate was chosen) and hierarchical modifications (presuming that a relation's constituent was not adequately placed

[2] This criterion states that a relation that holds between two textual spans holds also between the most salient units of the constituent spans.

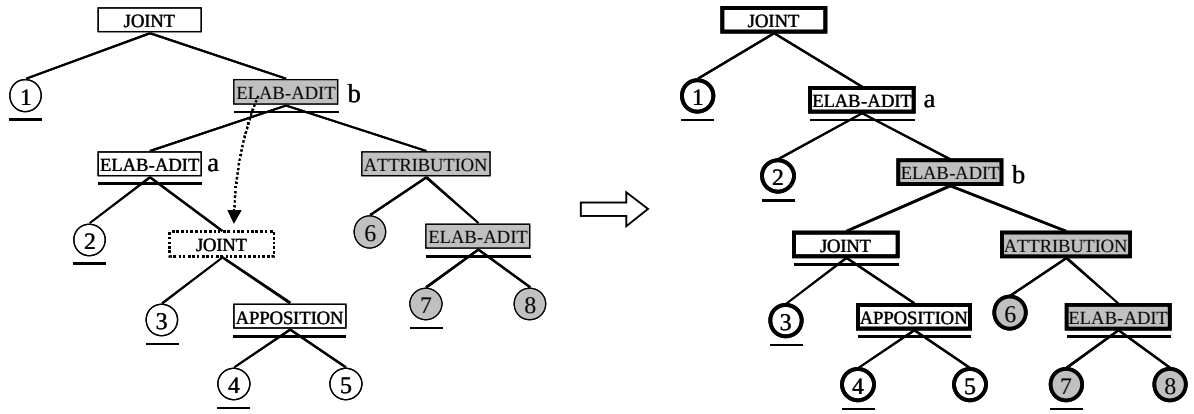

Figure 3 Hierarchical operation on the structure.

in the hierarchy). They both are aimed at counteracting the mistakes due to the mentioned uncertainties in one step of analysis.

While the first type of operation is quite simple, the second one is relatively complex and is similar to tree-adjoining (Joshi, 1987) operations. It involves two steps:

1. *cut:* a subtree is extracted from an internal frontier[3] of a node's daughter;
2. *paste:* the subtree obtained is adjoined as auxiliary tree to the opposite internal frontier of the other daughter.

The operations obey the *sequentiality principle* (Cristea and Webber, 1997); that is, a left to right reading of the terminal nodes of the structure corresponds to the initial ordering of units in text. That's why the extraction of the subtree in step one is done from the internal frontier (left, or right) of a daughter and its adjunction is done on a node from the opposite frontier (right or left, respectively) of the other daughter.

We illustrate this operation on the example presented in section 3.2. An operation on the structure hierarchy was performed in order to obtain the correct variant of structure from the initial one. Figure 3 shows how the subtree covering the units 6-8 was extracted from the right side of the problematic structure 2-8 and was adjoined to a node in the left side (the target of the adjunction is the node covering the span 3-5, as indicated by the arrow).

Observe that, actually, this kind of modification functions like an *undo* operation for the wrong adjunction of a partial tree to the already build structure, in a process of discourse parsing. In our example, there are multiple possibilities to adjunct structure 6-8 as auxiliary tree to the already build structure 1-5 (see Figure 3). Choosing as target of adjunction the node ELAB-ADIT(a) (2-5) will lead to the structure configuration on the left side, and choosing the node JOINT (3-5) will lead to the configuration on the right side. The modification we propose tries to counteract the wrong decision and therefore to make a different adjunction.

4.2 Local corrections

In the process of local correction, the basic operations are applied on a substructure of the whole discourse tree (that we call *working structure*), which is localized by the current exception and is identified as the common ancestor of the two units involved in the reference.

The process of correction of this substructure starts with verifying the correct assignation of nuclear roles and the correct associations of constituents, and then performs the necessary basic (nuclear or hierarchical) operations that allows the vein of anaphor's unit to pass through the antecedent's unit.

The verification of nuclear roles assignation is done first for the nodes on the path between the unit of antecedent and the root of the working structure, then on the path between the root and the unit of the anaphor.

[3] The *left (right) internal frontier* of a tree consists of the set of the leftmost (rightmost) non-terminal nodes from all depths in the tree.

If nuclear modifications are possible such that the first path becomes nuclear[4], the new vein will contain the antecedent's unit. Another successful structure configuration would be that in which the path antecedent-root (only up to the left daughter of the root) is nuclear and the path root-anaphor doesn't pass through a node that is right-satellite of another node (cf. the definition of the vein).
Otherwise, further verifications are done on the nodes from the internal frontiers of the two daughters of the working structure, and the valid association of constituents is questioned. The criterion followed is that of Marcu (1997), previously mentioned in Section 3.2.

4.3 Global corrections

The correction of the whole discourse structure allows local correction to be performed on the substructures determined by each exception, in a linear order given by the occurrence in the text of the anaphors involved. The local corrections are tried for each pair anaphor-antecedent, for all antecedents in that coreferences chain. A priority order is considered between the antecedents, given by the number of other existing references: the more a unit is referred to, the more likely to be a nucleus or to indicate an adjunction target.
The application of a local correction to the whole structure results in a substructure that is "better" and that contributes positively to the correctness of the whole one. We therefore claim that the successive application of local corrections is consistent, and the appropriate structure modifications do not give rise to other exceptions.

5 Implementation and evaluation

This section contains some details about the implementation and the functioning of the correction method, and it shows how the quality of the results is evaluated.

The material we used is a collection of 30 texts SGML annotated for coreferences[5] and for rhetorical structure[6].
In order to ease the processing of its tree-like rhetorical structure while updating the surface annotation, we decided to transform them into XML format and to apply modification scripts (written in JavaScript) on the document structure following the standard Document Object Model (DOM).
In a first step, we implemented the translation of the original annotation of rhetorical structure (that contains tags for *edus* and, separately, for the links between the structure nodes) into an annotation that reflects this structure on the surface text (it contains the tags that mark each node, and the relation parent-daughter is encoded by embedding tags).
The correction method, currently under implementation, is applied on the generated XML structure, as described in Section 4. The application runs as a script in a Web-browser, and the interaction with the user is assured by browser's dialog boxes. These features will allow us to develop it as a Web-application that will benefit of the advantages of platform- and software-independence.
In order to evaluate the correction method we intend to compare its results with the manually corrected version of the same corpus of texts.

6 Conclusion

This paper reports on on-going work towards the application of Veins Theory to correct discourse structures, based on referential constraints.
We proposed the basic correction operations on the tree-like structure of text, and local and global correction methods that can be applied in order to improve partial or complete structure annotations, either produced by humans or automatically derived (on the basis of cue-words for instance). Another possible application is that of guidance or assistance during an incremental discourse parsing process.
Related work, as that of (Schauer and Hahn, 2001), relies on the same idea of considering text cohesion in order to address the coherence

[4] We call *nuclear path* (in a tree-like structure) the path between two nodes that passes through nuclear nodes only.

[5] The MUC7 corpus (Hirschmann and Chinchor, 1997).

[6] By (Marcu et al., 1999).

problem of discourse. It proposes an algorithm for the combined computation of co-references and discourse structure, using the right frontier of the partially built tree to find both the target to connect a new unit to, and the antecedents of the anaphora in the new unit. Our approach is more specific than this with regard to the constraints imposed to the structure.

References

Cristea D. and Webber B.L. (1997) *Expectations in Incremental Discourse Processing*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid.

Cristea D., Ide N., and Romary L. (1998) *Veins Theory: A Model of Global Discourse Cohesion and Coherence*. In Proceedings of COLING-ACL'98, pp. 281–85.

Cristea D., Ide N., Marcu D., and Tablan M.V. (2000) *Discourse Structure and Co-Reference: An Empirical Study*. In Proceedings of the 18th International Conference on Computational Linguistics COLING'2000, Saarbrueken.

Fox B. (1987) *Discourse Structure and Anaphora. Written and Conversational English.* Cabbridge Studies in Linguistics, Cambridge University Press.

Grosz B., Joshi A., and Weinstein S. (1995) *Centering: A Framework for Modeling the Local Coherence of Discourse*. In Computational *Linguistics* 2(21), pp. 203-225.

Hirschman L. and Chinchor N. (1997) *MUC-7 Co-reference Task Definition*.

Ide N. and Cristea D. (2000) *A Hierarchical Account of Referential Accessibility*. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL'2000, Hong Kong.

Joshi A. (1987) *An introduction to Tree Adjoining Grammar*. In A. Manaster-Rammer (ed.) Mathematics of Language.

Knott A. and Dale R. (1994) *Using linguistic phenomena to motivate a set of coherence relations*. Discourse Processes, 18(1), 35-62.

Mann W.C. and Thompson S.A. (1988) Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text 8(3):243–281.*

Marcu D. (1997) *The rhetorical parsing, summarization and generation of natural language texts.* Ph. D. Thesis. Dept. of Computer Science, University of Toronto.

Marcu D. (1999) *Discourse trees are good indicators of importance in text*. In I. Mani and M. Maybury editors, Advances in Automatic Text Summarization, pages 123-136, The MIT Press.

Marcu D. (2000) *The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach.* Computational Linguistics, 26 (3), pp. 395-448.

Marcu D., Amorrortu E., and Romera M. (1999) *Experiments in Constructing a Corpus of Discourse Trees*. In ACL'99 Workshop on Standards and Tools for Discourse Tagging, Maryland.

Sanders T.J.M., Spooren W.P.M.S., and Noordman L.G.M. (1992) *Towards a taxonomy of coherence relations*. Discourse Processes, 15(1), 1-35.

Schauer H. and Hahn U. (2001). *Anaphoric Cues for Coherence Relations.* In Proceedings of RANLP'2001, pp. 228–235.