# Veins Theory:
# A Model of Global Discourse Cohesion and Coherence

Dan CRISTEA
Dept. of Computer Science
University «A.I. Cuza»
Iaşi, Romania
dcristea@infoiasi.ro

Nancy IDE
Dept. of Computer Science
Vassar College
Poughkeepsie, NY, USA
ide@cs.vassar.edu

Laurent ROMARY
Loria-CNRS
Vandoeuvre-Les-Nancy,
France
romary@loria.fr

## Abstract

In this paper, we propose a generalization of Centering Theory (CT) (Grosz, Joshi, Weinstein (1995)) called Veins Theory (VT), which extends the applicability of centering rules from local to global discourse. A key facet of the theory involves the identification of «veins» over discourse structure trees such as those defined in RST, which delimit domains of referential accessibility for each unit in a discourse. Once identified, reference chains can be extended across segment boundaries, thus enabling the application of CT over the entire discourse. We describe the processes by which veins are defined over discourse structure trees and how CT can be applied to global discourse by using these chains. We also define a discourse «smoothness» index which can be used to compare different discourse structures and interpretations, and show how VT can be used to abstract a span of text in the context of the whole discourse. Finally, we validate our theory by analyzing examples from corpora of English, French, and Romanian.

## Introduction

As originally postulated, Centering Theory (CT) (Grosz, Joshi, and Weinstein (1995)) accounts for references between adjacent units but is restricted to local reference (i.e., within segment boundaries). Recently, CT-based work has emerged which considers the relation of global discourse structure and anaphora, all of which proposes extensions to centering in order to apply it to global discourse.

We approach the relationship between global structure and anaphora resolution from a different, but related, perspective. We identify *domains of referential accessibility* for each discourse unit over discourse structure trees such as those defined in Rhetorical Structure Theory (RST ; Mann and Thompson (1987)) and show how CT can then be applied to global discourse by using these domains. As such, our approach differs from Walker's (1996), whose account of referentiality within the cache memory model does not rely on discourse structure, but rather on cue phrases and matching constraints together with constraints on the size of the cache imposed to reflect the plausible limits of the attentional span. Our approach is closer to that of Passonneau (1995) and Hahn and Strübe (1997), who both use a stack-based model of discourse structure based on Grosz and Sidner's (1986) focus spaces. Such a model is equivalent to a dynamic processing model of a tree-like structure reflecting the hierarchical nesting of discourse segments, and thus has significant similarities to discourse structure trees produced by RST (see Moser and Moore (1996)). However, using the RST notion of *nuclearity*, we go beyond previous work by revealing a "hidden" structure in the discourse tree, which we call *veins*, that enables us to determine the referential accessibility domain for each discourse unit and ultimately to apply CT globally, without extensions to CT or additional data structures.

In this paper, we describe *Veins Theory* (VT) by showing how veins are defined over discourse structure trees, and how CT can be applied to global discourse by using them. We use centering transitions (Brennan, Friedman and Pollard (1987)) to define a «smoothness» index, which is used to compare different discourse structures and interpretations. Because veins define the domains of referential access for each discourse unit, we further demonstrate how VT may be potentially used to determine the «minimal» parts of a text required to resolve references in a given utterance or, more generally, to understand it out of the context of the entire discourse. Finally, we validate our theory by analyzing examples from corpora of English, French, and Romanian.

## 1 The vein concept

We define *veins* over discourse structure trees of the kind used in RST. Following that theory, we consider the basic units of a discourse to be non-overlapping spans of text (i.e., sharing no common text), usually reduced to a clause and including a single predicate; and we assume that various rhetorical, cohesive, and coherence relations hold between individual units or groups of units. [1]

[1] Note that unlike RST, Veins Theory (VT) is not concerned with the type of relations which hold

We represent discourse structures as binary trees, where terminal nodes represent *discourse units* and non-terminal nodes represent *discourse relations.* A polarity is established among the children of a relation, which identifies at least one node, the *nucleus,* considered essential for the writer's purpose; non-nuclear nodes, which include spans of text that increase understanding but are not essential to the writer's purpose are called *satellites.*

Vein expressions defined over a discourse tree are sub-sequences of the sequence of units making up the discourse. In our discussion, the following notations are used:

- each terminal node (leaf node, discourse unit) has an attached label;
- *mark(x)* is a function that takes a string of symbols *x* and returns each symbol in x marked in some way (e.g., with parentheses);
- *simpl(x)* is a function that eliminates all marked symbols from its argument, if they exist; e.g. *simpl(a(bc)d(e))=ad*;
- *seq(x, y)* is a sequencing function that takes as input two non-intersecting strings of terminal node labels, *x* and *y*, and returns that permutation of *x*/*y* (*x* concatenated with *y*) that is given by the left to right reading of the sequence of labels in *x* and *y* on the terminal frontier of the tree. The function maintains the parentheses, if they exist, and *seq(nil, y) = y.*

**Heads**

1. The head of a terminal node is its label.
2. The head of a non-terminal node is the concatenation of the heads of its nuclear children.

**Vein expressions**

1. The vein expression of the root is its head.
2. For each nuclear node whose parent node has vein *v*, the vein expression is:
   - if the node has a left non-nuclear sibling with head *h*, then *seq(mark(h), v)*;
   - otherwise, *v*.
3. For each non-nuclear node of head *h* whose parent node has vein *v*, the vein expression is:
   - if the node is the left child of its parent, then *seq(h,v);*
   - otherwise, *seq(h, simpl(v)).*

Note that the computation of heads is bottom-up, while that of veins is top-down.

Consider example 1:

```
1.  According  to engineering  lore,
2.  the late Ermal C. Fraze,
3.  founder of Dayton Reliable  Tool &
    Manufacturing Company in Ohio,
2a. came up with a practical idea for
    the pop-top lid
3. after    attempting    with    halting
   success  to  open  a  beer  can  on  the
   bumper of his car.
```

The structure of this discourse fragment is given in Figure 1. The central gray line traces the principal vein of the tree, which starts at the root and descends along the nuclear nodes. Auxiliary veins are attached to the principal vein. The vein expressions corresponding to each node indicate its domain of accessibility, as defined in the following section. Accordingly, in this example, unit 1 is accessible from unit 2, but not unit 3.

## 2 Accessibility

The domain of accessibility of a unit is defined as the string of units appearing in its vein expression and prefixing that unit itself.

More formally, for each terminal node *u*, if *vein(u)* is its vein, then accessibility from *u* is given by *acc(u) = pref(u, unmark(vein(u))),* where:

- *vein* is the function that computes the vein;
- *unmark(x)* is a function that removes the markers from all symbols of its argument;
- *pref* is a function that retains the prefix of the second argument up to and including the first argument (e.g., if $\alpha$ and $\beta$ are strings of labels and *u* is a label, $pref(u, \alpha u\beta)=\alpha u$.

**Conjecture C1: References from a given unit are possible only in its domain of accessibility.**

In particular, we can say the following:

1. In most cases, if *B* is a unit and $b \in B$ is a referential expression, then either *b* directly realizes a center that appears for the first time in the discourse, or it refers back to another center realized by a referential expression $a \in A$, such that $A \in acc(B)$.[2] Such cases instantiate *direct references.*
2. If (1) is not applicable, then if *A, B,* and *C* are units, $c \in C$ is a referential expression that refers to $b \in B$, and *B* is not on the vein of *C* (i.e., it is not visible from *C*), then there is an item $a \in A$, where *A* is a unit on the common vein of *B* and *C*, such that both *b* and *c* refer to *a*. In this case we say that *c* is an *indirect reference* to *a*.[3]
3. If neither (1) nor (2) is applicable, then the reference in *C* can be understood without the referee, as if the corresponding entity were introduced in the discourse for the first time. Such references are *inferential references.*

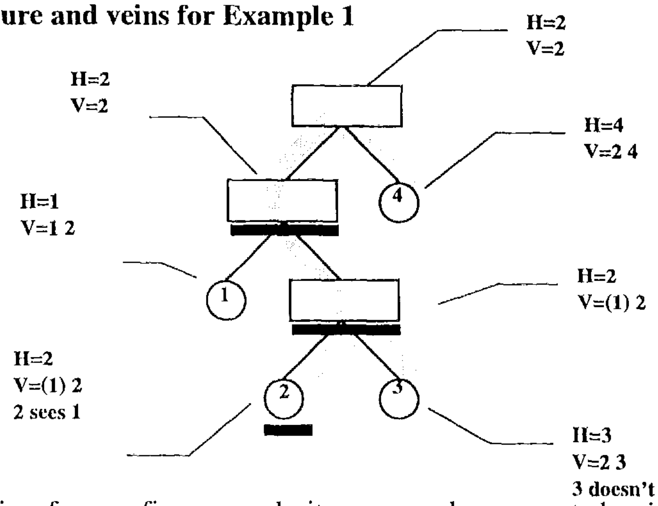
Note that VT is applicable even when the division into units is coarser than in our examples. For instance, Example 1 in its entirety could be taken to comprise a single unit; if it appeared in the context of a larger discourse, it would still be possible to compute its veins (although, of course, the veins would likely be shorter because there are fewer units to consider). It can be proven formally (Cristea,

among discourse units, but considers only the topological structure and the nuclear/satellite status (see below) of discourse units.

[2] If *a* and *b* are referential expressions, where the center (directly) realized by *b* is the same as the one (directly) realized by *a*, or where it is a role of the center (directly) realized by *a*, we will say that *b* refers (back) to *a*, or *b* is a bridge reference to *a*.

[3] On the basis of their common semantic representations.

Figure 1: Tree structure and veins for Example 1



1998) that when passing from a finer granularity to a coarser one the accessibility constraints are still obeyed. This observation is important in relation to other approaches that search for stability with respect to granularity (see for instance, Walker, 1996).

## 3 Global coherence

This section shows how VT can predict the inference load for processing global discourse, thus providing an account of discourse coherence.

A corollary of Conjecture C1 is that CT can be applied along the accessibility domains defined by the veins of the discourse structure, rather than to sequentially placed units within a single discourse segment. Therefore, in VT reference domains for any node may include units that are sequentially distant in the text stream, and thus long-distance references (including those requiring "return-pops" (Fox, 1987) over segments that contain syntactically feasible referents) can be accounted for. Thus our model provides a description of *global discourse cohesion*, which significantly extends the model of local cohesion provided by CT.

CT defines a set of transition types for discourse (Grosz, Joshi, and Weinstein (1995); Brennan, Friedman and Pollard (1987)). A *smoothness score* for a discourse segment can be computed by attaching an elementary score to each transition between sequential units according to Table 2, summing up the scores for each transition in the entire segment, and dividing the result by the number of transitions in the segment. This provides an index of the overall coherence of the segment.

A *global CT smoothness score* can be computed by adding up the scores for the sequence of units making up the whole discourse, and dividing the result by the total number of transitions (number of units minus one). In general, this score will be slightly higher than the average of the scores for the individual segments, since accidental transitions at segment boundaries might also occur. Analogously, a *global VT smoothness score* can be computed using accessibility domains to determine transitions rather than sequential units.

Table 2: Smoothness scores for transitions

| | |
|---|---|
| CENTER CONTINUATION | 4 |
| CENTER RETAINING | 3 |
| CENTER SHIFTING (SMOOTH) | 2 |
| CENTER SHIFTING (ABRUPT) | 1 |
| NO Cb | 0 |

**Conjecture C2: The global smoothness score of a discourse when computed following VT is at least as high as the score computed following CT.**

That is, we claim that long-distance transitions computed using VT are systematically smoother than accidental transitions at segment boundaries. Note that this conjecture is consistent with results reported by authors like Passonneau (1995) and Walker (1996), and provides an explanation for their results.

We can also consider anaphora resolution using Cb's computed using accessibility domains. Because a unit can simultaneously occur in several accessibility domains, unification can be applied using the Cf list of one unit and those of possibly several subsequent (although not necessarily adjacent) units. A graph of Cb-unifications can be derived, in which each edge of the graph represents a Cb computation and therefore a unification process.

## 4 Minimal text

The notion that text summaries can be created by extracting the nuclei from RST trees is well known in the literature (Mann and Thompson, (1988)). Most recently, Marcu (1997) has described a method for text summarization based on nuclearity and selective retention of hierarchical fragments. Because his *salient units* correspond to *heads* in VT, his results are predicted in our model. That is, the union of heads at a given level in the tree provides a summary of the text at a degree of detail dependent on the depth of that level.

In addition to summarizing entire texts, VT can be used to summarize a given unit or sub-tree of that text. In effect, we reverse the problem addressed by text summarization efforts so far: instead of attempting to summarize an entire discourse at a given level of detail, we select a single span of text and abstract the minimal text required to understand *this span alone* when considered in the context of the entire discourse. This provides a kind of *focused abstraction,* enabling the extraction of sub-texts from larger documents. Because vein expressions for each node include all of the nodes in the discourse within its domain of reference, they identify exactly which parts of the discourse tree are required in order to understand and resolve references for the unit or subtree below that node.

**Table 5: Verifying conjecture C1**

| Source | No. of units | Total no. of refs | Direct on the vein (case 1) | | Indirect on the vein (case 2) | | Inference (case 3) | | How many obey C1 | |
|---|---|---|---|---|---|---|---|---|---|---|
| English | 62 | 97 | 75 | 77.3% | 14 | 14.4% | 5 | 5.2% | 94 | 96.9% |
| French | 48 | 110 | 98 | 89.1% | 11 | 10.0% | 1 | 0.9% | 110 | 100.0% |
| Romanian | 66 | 111 | 104 | 93.7% | 2 | 1.8% | 5 | 4.5% | 111 | 100.0% |
| **Total** | **176** | **318** | **277** | **87.1%** | **27** | **8.5%** | **11** | **3.5%** | **315** | **99.1%** |

**Table 6: Verifying Conjecture C2**

| Source | No. of transitions | CT Score | Average CT score per transition | VT score | Average VT score per transition |
|---|---|---|---|---|---|
| English | 59 | 76 | 1.25 | 84 | 1.38 |
| French | 47 | 109 | 2.32 | 116 | 2.47 |
| Romanian | 65 | 142 | 2.18 | 152 | 2.34 |
| **Total** | **173** | **327** | **1.89** | **352** | **2.03** |

## 5. Corpus analysis

Because of the lack of large-scale corpora annotated for discourse, our study currently involves only a small corpus of English, Romanian, and French texts. The corpus was prepared using an encoding scheme for discourse structure (Cristea, Ide, and Romary, 1998) based on the Corpus Encoding Standard (CES) (Ide (1998)). The following texts were included in our analysis:

- three short English texts, RST-analyzed by experts and subsequently annotated for reference and Cf lists by the authors;
- a fragment from de Balzac's «Le Père Goriot» (French), previously annotated for co-reference (Bruneseaux and Romary (1997)); RST and Cf lists annotation made by the authors;
- a fragment from Alexandru Mitru's «Legendele Olimpului»[4] (Romanian); structure, reference, and Cf lists annotated by one of the authors.

The encoding marks referring expressions, links between referring expressions (co-reference or functional), units, relations between units (if known), nuclearity, and the units' Cf lists in terms of referring expressions. We have developed a program[5] that does the following: builds the tree structure of units and relations between them, adds to each referring expression the index of the unit it occurs in, computes the heads and veins for all nodes in the structure, determines the accessibility domains of the terminal nodes (units), counts the number of direct and indirect references.

Hand-analysis was then applied to determine which references are inferential and therefore do not conform to Conjecture C1, as summarized in Table 5. Among the 318 references in the text, only three references not conforming to Conjecture C1 were found (all of them appear in one of the English texts). However, if the BACKGROUND relation is treated as bi-nuclear,[6] all three of these references become direct.

To verify Conjecture C2, Cb's and transitions were first marked following the sequential order of the units (according to classical CT), and a smoothness score was computed. Then, following VT, accessibility domains were used to determine maximal chains of accessibility strings, Cb's and transitions were re-computed following these strings, and a VT smoothness score was similarly computed. The results are summarized in Table 6. They show that the score for VT is better than that for CT in all cases, thus validating Conjecture C2.

An investigation of the number of long-distance resolutions yielded the results shown in Table 7. Such resolutions could not have been predicted using CT.

**Table 7: Long distance reference resolution**

| Source | No of long distance Cb unifications | No of new referents found |
|---|---|---|
| English | 6 | 2 |
| French | 11 | 1 |
| Romanian | 18 | 3 |

[4] «The Legends of Olimp»

[5] Written in Java.

[6] Other bi-nuclear relations are JOIN and SEQUENCE.

## 6. Discussion and related work

VT is not a model of anaphora resolution; rather, its accessibility domains provide a means to constrain the resolution of anaphora. The fundamental assumption underlying VT is that an inter-unit reference is possible *only if the two units are in a structural relation with one another,* even if they are distant from one another in the text stream. Furthermore, inter-unit-references are *primarily to nuclei rather than to satellites,* reflecting the intuition that nuclei assert the writer's main ideas and provide the main «threads» of the discourse (Mann and Thompson [1988]. This is shown in the computation of veins over (binary) discourse trees where each pair of descendants of a parent node are either both nuclear or the nuclear node is on the left (a *left-polarized tree).* In such trees, any reference from a nuclear unit must be to entities contained in linguistic expressions appearing in previously occurring nuclei (although perhaps not *any* nucleus). On the other hand, satellites are dependent on their nuclei for their meaning and hence may refer to entities introduced within them. The definition of veins formalizes these relationships. Given the mapping of Grosz and Sidner's (1986) stack-based model of discourse structure to RST structure trees outlined by Moser and Moore (1996), the domains of referentiality defined for left-polarized trees using VT are consistent with those defined using the stack-based model (e.g. Passonneau (1995), Hahn and Strübe (1997)). However, in cases where the discourse structure is not left-polarized, VT provides a more natural account of referential accessibility than the stack-based model. In non left-polarized trees, at least one satellite precedes its nucleus in the discourse and is therefore its left sibling in the binary discourse tree. The vein definition formalizes the intuition that in a sequence of units A B C, where A and C are satellites of B, B can refer to entities in A (its left satellite), but the subsequent right satellite, C, cannot refer to A due to the interposition of nuclear unit B. In stack-based approaches to referentiality, such configurations pose problems: because B dominates[7] A it must appear below it on the stack, even though it is processed after A. Even if the processing difficulties are overcome, this situation leads to the postulation of cataphoric references when a satellite precedes its nucleus, which is counter-intuitive.

## Acknowledgements

Our thanks go to Daniel Marcu who pointed some weak parts and provided RST analysis and to the TELRI program who facilitated the second meeting of the three authors.

[7] We use Grosz and Sidner's (1986) terminology here, but note the equivalence of dominance in G&S and nucleus/satellite relations in RST pointed out by Moser and Moore (1996).

## References

Brennan, S.E., Walker Friedman, M. and Pollard, C.J. (1987). A Centering Approach to Pronouns. *Proceedings of the 25th Annual Meeting of the ACL,* Stanford, 155-162.

Bruneseaux Florence and Laurent Romary (1997). Codage des Références et coréférences dans les Dialogues Homme-machine. *ProceeCings ofACH/ALLC,* Kingston (Ontario).

Cristea, D. (1998). Formal proofs in Incremental Discourse Processing and Veins Theory, Research Report TR98-2 Dept. of Computer Science, University "A.I.Cuza", Iaşi.

Cristea, D., Ide, N. and Romary, L. (1998). Marking-up Multiple Views of a Text: Discourse and Reference, *Proceedings of the First International Conference on Language Resources and Evaluation,* Granada, Spain.

Fox, B. (1987). Discourse Structure and Anaphora. Written and Conversational English. no 48 in Cambridge Studies in Linguistics, Cambridge University Press.

Grosz, B.J., Joshi, A.K. and Weinstein, S. (1995) Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics,* 12(2), 203-225.

Grosz, B. and Sidner, C. (1986). Attention, Intention and the Structure of Discourse. *Computational Linguistics,* 12, 175-204.

Hahn, U. and Strübe, M. (1997). Centered Segmentation: Scaling Up the Centering Model to Global Discourse Sttructure. *Proceedings of EACL/ACL'97,* Madrid, 104-11.

Ide, N. (1998) Corpus Encoding Standard: Encoding Practices and a Data Aarchitecture for Linguistic Corpora. *Proceedings of the First International Conference on Language Resources and Evaluation,* Granada, Spain. See also http://www.cs.vassar.edu/CES/.

Mann, W.C., Thompson S.A. (1988). Rhetorical structure theory: A theory of text organization, *Text,* 8:3, 243-281.

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

Moser, M. and Moore, J. (1996). Toward a Synthesis of Two Accounts of Discourse Structure. *Computational Linguistics,* 22:3, 409-20.

Passonneau, R.J. (1995). Using Centering to Relax Gricean Informational Constraints on Discourse Anaphoric Noun Phrases, research report, Bellcore.

Walker, M.A. (1996). The Cash Memory Model. *Computational Linguistics,*22:2, 255-64.

Walker, M.A.; Joshi, A.K., Prince, E.F. (1997). Centering in Naturally-Occurring Discourse: An Overview. In Walker, M.A.; Joshi, A.K., Prince, E.F. (eds.): *Centering in Discourse,* Oxford University Press.