Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

# HEURÍSTICAS DE SUMARIZAÇÃO DE ESTRUTURAS RST

Eloize Rossi Marques Seno
Lucia Helena Machado Rino

**NILC-TR-05-04**

Fevereiro, 2005

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

# Resumo

O propósito deste relatório é apresentar um elenco de heurísticas de poda de estruturas retóricas para a sumarização automática de textos representados por suas subjacentes de estruturas retóricas, construídas segundo a *Rhetorical Structure Theory – RST*.

Com base em características específicas das relações da RST, as heurísticas de poda visam identificar e excluir as informações menos relevantes da estrutura RST de um texto sem, contudo, deixar de preservar a coerência do discurso. Neste trabalho, a única questão contemplada é a quebra de coerência introduzida pelo tratamento inadequado das cadeias de co-referências (CCRs).

Para garantir que as CCRs sejam tratadas adequadamente, as heurísticas se baseiam em restrições impostas pela *Veins Theory - VT*, que delimita o domínio de acessibilidade referencial para cada unidade do discurso na forma de "veias", determinando os limites nos quais os antecedentes de uma anáfora podem ocorrer ao longo do discurso.

O elenco de heurísticas de poda compõe o protótipo de um sumarizador automático de estruturas RST, ainda em desenvolvimento.



Desenvolvimento para o Futuro
Departamento de Computação

ufscar

# Índice Geral

# Índice de Figuras

# Índice de Tabelas

## 1. Introdução

Este relatório descreve um elenco de heurísticas de poda para a sumarização automática de estruturas RST, que são estruturas retóricas de textos construídas com base na *Rhetorical Structure Theory – RST* (Mann and Thompson, 1987). A estrutura RST de um texto representa o inter-relacionamento entre suas unidades discursivas. Por hipótese, se o texto correspondente for coerente, sua estrutura RST permitirá recuperar sua mensagem por meio das relações de significado entre suas unidades discursivas. Neste caso, a estrutura RST permite identificar as informações supérfluas do texto subjacente, para exclusão durante a construção do seu sumário.

A identificação das informações irrelevantes não é o único problema da sumarização automática: ao excluí-las da estrutura RST original, esta deve ser inteiramente reestruturada, para garantir a preservação da coerência do sumário correspondente. Um dos problemas mais sérios de quebra de coerência, neste caso, é a ocorrência de quebras de cadeias de co-referências (CCRs) – referências a um objeto já introduzido na comunicação, reproduzidas ao longo do discurso (Paraboni, 1997), que ocorre quando a unidade discursiva que contém um termo anafórico é inclusa no sumário, mas a unidade discursiva que contém o seu antecedente não.

É válido ressaltar que a RST não impõe nenhuma restrição para preservar o relacionamento co-referencial entre as unidades discursivas e, portanto, os sumários produzidos com base nessa abordagem podem apresentar problemas de co-referenciação. A *Veins Theory* - *VT* (Cristea et al., 1998) propõe contornar esse problema, delimitando domínios de acessibilidade referencial para cada unidade do discurso na forma de "veias" definidas sobre a estrutura RST. Tais veias determinam os limites nos quais os antecedentes de um termo anafórico podem ocorrer ao longo do discurso. Assim, para manter a coerência é necessário que a CCR completa esteja contida numa única veia (ou seja, tanto o antecedente quanto o termo anafórico devem estar na mesma veia).

As heurísticas propostas neste trabalho, baseadas tanto na RST quanto na *Veins Theory*, visam agregar o potencial de estruturação da RST à preservação da coerência potencializada pela *Veins Theory*. Assim, as heurísticas procuram informações supérfluas na estrutura RST de um texto guiadas por dois aspectos principais dessas teorias: as características específicas das relações retóricas da RST e o domínio de acessibilidade referencial da *Veins Theory*.

As seções 2 e 3 descrevem brevemente a RST e a *Veins Theory*, respectivamente. A seção 4 apresenta as heurísticas de poda, seguindo-se a proposta de sua aplicação para a sumarização automática de estruturas RST, na seção 5. Por fim, a seção 6 apresenta algumas considerações finais.

## 2. Rhetorical Structure Theory

A RST (Mann and Thompson, 1987) fundamenta-se no princípio de que um texto tem uma estrutura retórica subjacente e que, através dessa estrutura, é possível recuperar o objetivo comunicativo que o escritor do texto pretendeu atingir ao escrevê-lo. Essa estrutura é composta por unidades elementares do discurso (*Elementary Discourse Unit* ou *EDUs*, no inglês), inter-relacionadas por meio de relações retóricas. As *EDUs* são unidades mínimas de significado que compõem um texto. As relações retóricas indicam os tipos de relações existentes entre tais unidades, visando a organização coerente de um texto ou discurso.

Segundo a RST, as relações retóricas inter-relacionam *EDUs* que são expressas por segmentos adjacentes em um texto. A cada *EDU* é atribuído um papel de núcleo ou satélite. O núcleo, ou unidade nuclear, expressa a informação principal, sendo, portanto, mais relevante do que o satélite. O satélite apresenta informação adicional, a qual exerce influência na interpretação do leitor sobre a informação apresentada no núcleo. Assim, núcleos, na maioria das vezes, são compreensíveis independentemente dos satélites, mas não vice-versa. Vale ressaltar que, embora na teoria isso possa ser verdade, na prática muitas vezes torna-se impossível à compreensão do núcleo sem o seu satélite, casos em que os satélites são essenciais para manter a coerência e garantir o fluxo normal do discurso.

Há casos em que ambas as unidades são nucleares, ou seja, ambas apresentam informações importantes. Assim, as relações RST são divididas em duas classes: hipotáticas e paratáticas (Marcu, 1997). As relações hipotáticas inter-relacionam pares de *EDUs* que apresentam diferentes graus de importância, sendo uma nuclear e a outra satélite. Essas relações denominam-se mononucleares. As relações paratáticas inter-relacionam *EDUs* que apresentam o mesmo grau de importância e são denominadas relações multinucleares.

Segundo Mann e Thompson, o conjunto de relações retóricas da RST é capaz de representar todas as possíveis relações de significado entre os segmentos discursivos de uma grande gama de textos. Esse conjunto de relações é mostrado na Tabela 1[1].

Tabela 1: Conjunto de relações retóricas

| **Relação Retórica** | **Tipo de Relação** |
|---|---|
| ANTITHESIS | Mononuclear |
| BACKGROUND | Mononuclear |
| CIRCUMSTANCE | Mononuclear |
| CONCESSION | Mononuclear |
| CONDITION | Mononuclear |
| CONTRAST | Multinuclear |
| ELABORATION | Mononuclear |
| ENABLEMENT | Mononuclear |
| EVALUATION | Mononuclear |
| EVIDENCE | Mononuclear |
| INTERPRETATION | Mononuclear |
| JOINT | Multinuclear |
| JUSTIFY | Mononuclear |
| MOTIVATION | Mononuclear |
| NON-VOLITIONAL CAUSE | Mononuclear |
| NON-VOLITIONAL RESULT | Mononuclear |
| OTHERWISE | Mononuclear |
| PURPOSE | Mononuclear |
| RESTATEMENT | Mononuclear |
| SEQUENCE | Multinuclear |
| SOLUTIONHOOD | Mononuclear |
| SUMMARY | Mononuclear |
| VOLITIONAL CAUSE | Mononuclear |
| VOLITIONAL RESULT | Mononuclear |

[1] Mantidas com a nomenclatura original, suas definições podem ser recuperadas da obra de referência.

Considere, por exemplo, as Figura 1 e 2, as quais ilustram relações mononuclear e multinuclear, respectivamente. No texto da Figura 1[2] a *EDU* 1 é o núcleo (N) e a *EDU* 2 é o satélite (S) da relação retórica *PURPOSE*, cujo satélite apresenta uma situação que será realizada através da atividade apresentada no núcleo. Em outras palavras pode-se dizer que uma das possíveis interpretações dessa estrutura seria N a fim de realizar S, como ilustra o texto correspondente. No texto da Figura 2 há uma relação retórica *SEQUENCE* indicando a seqüência de eventos entre as *EDUs* 1 e 2, sendo que as duas possuem o mesmo grau de importância. Como mostra a Tabela 1, o conjunto de relações da RST inclui apenas três relações multinucleares.

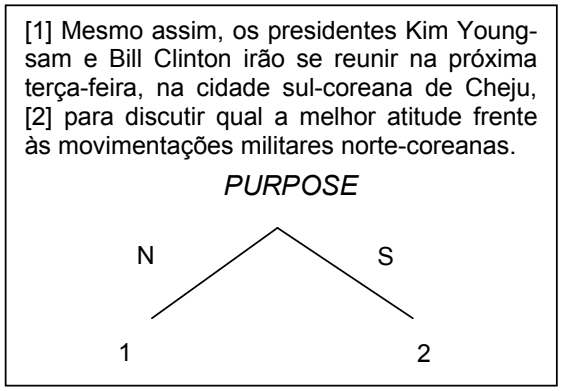


Figura 1: Exemplo de relação mononuclear

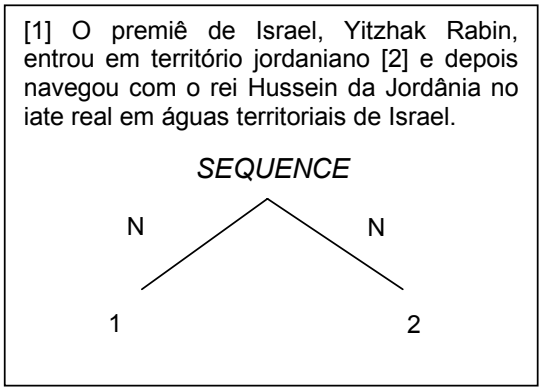


Figura 2: Exemplo de relação multinuclear

A definição de cada relação retórica consiste de quatro tipos de informações que o analista de um texto deve considerar, para determinar como duas *EDUs* se inter-relacionam. São elas:

- Restrições sobre o núcleo (N);
- Restrições sobre o satélite (S);
- Restrições sobre a combinação do núcleo e do satélite (N+S);
- Efeito (ou intenção do escritor): especifica o efeito que a relação causa no leitor ao interpretar o texto, ou o efeito pretendido pelo escritor ao selecionar tal relação para estruturar seu texto.

Para exemplificar, a Tabela 2 apresenta as definições das relações retóricas *ELABORATION, EVIDENCE, JUSTIFY* e *SEQUENCE*.

[2] Neste relatório, todos os textos-exemplo foram extraídos do corpus TeMário (Pardo e Rino, 2003) e apresentam as *EDUs* numeradas, para referência.

Tabela 2: Exemplos de definição de relações retóricas

| | |
|---|---|
| **Nome da relação:** *ELABORATION* | |
| **Restrições sobre N** | Não tem restrições |
| **Restrições sobre S** | Não tem restrições |
| **Restrições sobre N+S** | S apresenta detalhes adicionais sobre a situação ou algum elemento apresentado em N |
| **Efeito** | Leitor reconhece que S apresenta detalhes adicionais sobre N |
| **Nome da relação:** *EVIDENCE* | |
| **Restrições sobre N** | Leitor pode não acreditar em N com o grau de satisfação esperado pelo escritor |
| **Restrições sobre S** | Leitor acredita em S ou poderá acreditar facilmente em S |
| **Restrições sobre N+S** | A compreensão do leitor em S aumenta sua crença em N |
| **Efeito** | Leitor aumenta sua crença na asserção apresentada em N |
| **Nome da relação:** *JUSTIFY* | |
| **Restrições sobre N** | Não tem restrições |
| **Restrições sobre S** | Não tem restrições |
| **Restrições sobre N+S** | Escritor acredita que a compreensão do leitor em S aumenta sua disposição para aceitar a asserção apresentada em N |
| **Efeito** | A disposição do leitor para aceitar a asserção apresentada em N é aumentada |
| **Nome da relação:** *SEQUENCE* | |
| **Restrições sobre N** | Multinuclear |
| **Restrições sobre os N** | Sucessão de acontecimentos entre as situações presentes nos N |
| **Efeito** | Leitor reconhece a sucessão de acontecimentos presentes nos N |

Pode-se extrair a estrutura discursiva de um texto reconhecendo-se relações individuais entre pares de segmentos (ou entre múltiplos segmentos, no caso das relações multinucleares). A construção da estrutura RST completa é composicional: relações retóricas entre unidades proposicionais elementares são indicadas por orações simples, por exemplo. Essas sub-estruturas, por sua vez, podem compor estruturas mais complexas, as quais podem ser formadas por segmentos textuais mais elaborados (sentenças compostas ou justapostas, por exemplo). Assim, a estrutura discursiva pode ser representada por uma árvore retórica, cujos nós folha correspondem às *EDUs* e cujos nós internos representam relações retóricas, como mostram as Figuras 1 e 2.

Ao desenvolver o primeiro analisador retórico automático para o inglês, Marcu (1997, 2000) introduziu novas relações retóricas na RST para tratar um fenômeno lingüístico particular: os segmentos encaixados como, por exemplo, as orações subordinadas. Assim, ele modificou esse conjunto, introduzindo relações retóricas encaixadas (*embedded*, no inglês), derivando as relações indicadas por "-e" no final de seu nome.

Considere o exemplo da Figura 3, na qual a *EDU* 2 é o segmento encaixado, aqui relacionado à *EDU* 1 pela relação *ELABORATION-e*. Para representar esses casos,

em que há segmentos encaixados constituindo uma única unidade elementar, faz-se uso da relação multinuclear *SAME-UNIT* também proposta por Marcu (1997).

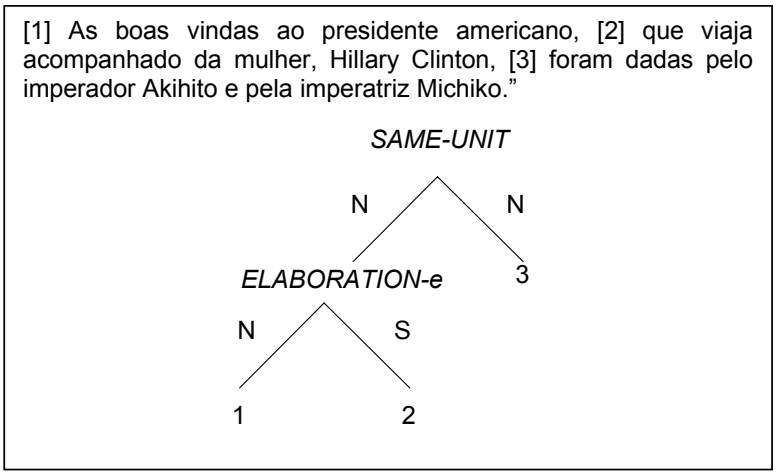


Figura 3: Estrutura retórica com segmento encaixado

Para a sumarização automática, dois fatores tornam a RST interessante: a) a nuclearidade pode corresponder à relevância, ou seja, uma unidade nuclear pode fornecer informação mais relevante do que o seu satélite e b) a escolha das unidades discursivas para a composição dos sumários pode basear-se na nuclearidade. No entanto, há situações em que satélites não podem ser desprezados, ou porque servem para complementar as informações nucleares e, assim, detalhar melhor a mensagem subjacente, ou para garantir que o discurso resultante seja coerente. A identificação dessas situações se baseia na função retórica do discurso em foco, razão pela qual a sumarização automática pode ser modelada a partir da RST (Sparck Jones, 1993; Ono et al., 1994; Marcu, 1997, 1999, 2000; O'Donnel, 1997).

## 3. Veins Theory

A *Veins Theory – VT* (Cristea et al., 1998) é uma generalização da *Centering Theory – CT* (Grosz et al., 1995), que trata da coerência local do discurso. A *Veins Theory* expande as regras de coerência local da *Centering Theory* para abranger a composicionalidade das unidades do discurso. Assim, visa garantir que um discurso todo seja coerente a partir da garantia da coerência local. Tais regras propõem um relacionamento entre a estrutura RST de um texto e suas cadeias de co-referências (CCRs). A co-referência pode ser estabelecida por meio de pronomes pessoais (por ex., "ela"), pronomes possessivos (por ex., "seu"), descrições definidas (por ex., "a mãe de Jõao"), dentre outras formas lingüísticas. A forma mais comum de co-referência é chamada relação anafórica, estabelecida entre uma expressão de referência (o termo anafórico) e um termo referente que a antecede no discurso (o termo antecedente), no qual ambos se referem à mesma entidade no discurso. Para efeito de ilustração, considere o trecho de texto apresentado a seguir, no qual o termo antecedente e o termo anafórico são apresentados em negrito.

"[1] **A Febraban (Federação Brasileira das Associações dos Bancos)** defende um regime básico com benefícios de até dois salários mínimos... [2] A proposta d**a**

**Febraban** prevê o financiamento através de contribuição dos segurados e das empresas até a faixa de dois salários mínimos".

Com base na noção de nuclearidade da RST, a *Veins Theory* delimita o domínio de acessibilidade referencial para cada unidade do discurso na forma de "veias" sobre a estrutura RST. A veia de uma unidade é definida como um conjunto de unidades do discurso que contêm o antecedente de um termo anafórico, pertencente àquela unidade. Tais veias determinam os limites nos quais os antecedentes de um termo anafórico podem ocorrer ao longo de um discurso coerente. Assim, a referência de uma anáfora ao seu antecedente deve ser resolvida na veia da própria unidade que contém a anáfora[3].

As veias definidas sobre uma árvore RST são subseqüências da seqüência de unidades que compõem o discurso e são computadas como segue (Cristea et al., 1998):

*Para todo n ∈ ARST*
*Se n é um nó folha*
*então head de n é igual a n*
*Senão*
*head de n é igual à concatenação dos heads dos seus filhos nucleares*
*Se n é o núcleo raiz da ARST, isto é, o núcleo mais nuclear*
*então veia de n é igual ao seu head*
*Para todo n núcleo cujo n pai tem uma veia v*
*Se n tem um irmão satélite à sua esquerda com head h*
*então veia de n é igual a seq(mark(h), v)*
*Senão*
*veia de n é igual a v*
*Para todo n satélite de head h cujo n pai tem uma veia v*
*Se n é o filho esquerdo do seu n pai*
*então veia de n é igual a seq(h,v)*
*Senão*
*veia de n é igual a seq(h, simpl(v))*

para:

*ARST: árvore RST de um texto-fonte qualquer;*
*n: nó da ARST em foco;*
*head de n: conjunto de unidades mais salientes de n, isto é, as unidades mais importantes no segmento de discurso correspondente;*
*mark(x): função que dada uma string de símbolos x, retorna cada símbolo em x marcado de alguma forma (por exemplo, com parênteses ou colchetes);*
*simpl(x): função que elimina todos os símbolos marcados dos seus argumentos (se existir algum), por exemplo, simpl(a(bc)d(e)) retorna ad;*
*seq(x, y): função que pega como entrada duas strings não-intersectadas de nós folhas, x e y, e retorna a permutação de x concatenado a y, dada pela seqüência de leitura de x e y na ARST.*

A objetivo desse algoritmo é determinar para cada nó da árvore RST de um texto a sua veia, isto é, o conjunto de nós que pode conter os antecedentes das anáforas pertencentes àquelas unidades.

[3] A resolução de uma anáfora consiste em determinar o antecedente correto para o qual o termo anafórico se refere.

Considere, por exemplo, o texto-exemplo da Figura 4 e sua correspondente árvore RST, ilustrada na Figura 5. Aplicando-se o cômputo das veias para cada nó da árvore obtêm-se os seguintes *heads* (*h*) e veias (*v*), apresentados em itálico nessa figura: h= 1 e v= 1, 3 para o nó 1; h= 2 e v= 1, 2, 3 para o nó 2; h= 3 e v= 1, 3 para o nó 3; h= 4 e v= 1, 3, 4 para o nó 4; h= 5 e v= 1, 3, 4, 5 para o nó 5; h= 6 e v= 1, 3, 6 para o nó 6; h= 7 e v= 1, 3, 7, 8 para o nó 7; h= 8 e v= 1, 3, (7), 8 para o nó 8; h= 9 e v= 1, 3, 8, 9, 10 para o nó 9; h= 10 e v= 1, 3, 8, 9, 10 para o nó 10.

[1] A empresa Produtos Pirata Indústria e Comércio Ltda., de Contagem [2] (na região metropolitana de Belo Horizonte), [3] deverá registrar este ano um crescimento de produtividade nas suas áreas comercial e industrial de 11% e 17%, respectivamente. [4] Os ganhos são atribuídos pela diretoria da fábrica à nova filosofia que vem sendo implantada na empresa desde outubro do ano passado, [5] quando a Pirata se iniciou no Programa Sebrae de Qualidade Total.

[6] Dona de 65% do mercado mineiro de temperos, condimentos e molhos, a Pirata reúne atualmente 220 funcionários. [7] A coordenadora do programa de qualidade na empresa, Márcia Cristina de Oliveira Neto, disse que [8] ainda não é possível dimensionar os ganhos financeiros que "certamente" a empresa terá, em conseqüência da melhoria da qualidade de seus produtos e serviços. [9] Por enquanto, os benefícios mais visíveis, segundo ela, são a organização e a limpeza da fábrica. [10] "Também a relação entre as pessoas tem melhorado bastante. As informações estão mais claras e os funcionários e clientes, mais satisfeitos".

Figura 4: Texto-exemplo

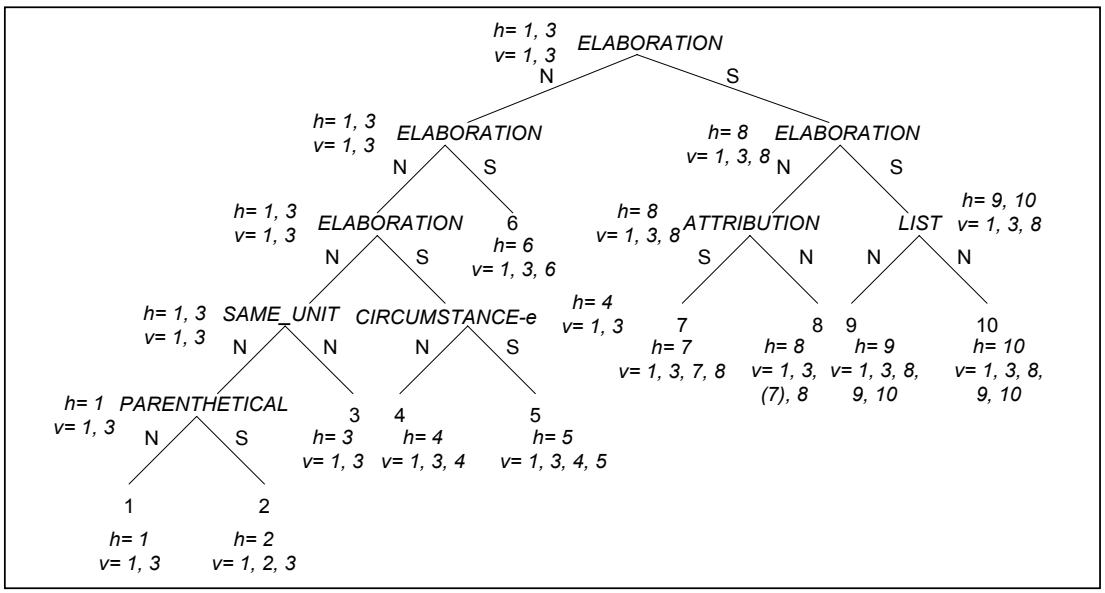


Figura 5: Cômputo das veias para a árvore RST do texto-exemplo

Para melhor ilustrar o domínio de acessibilidade referencial, considere, por exemplo, a referência anafórica (em negrito) no segmento de discurso correspondente ao nó 6, reproduzido a seguir.

> [6] Dona de 65% do mercado mineiro de temperos, condimentos e molhos, **a Pirata** reúne atualmente 220 funcionários.

Conforme a Figura 5, o domínio de acessibilidade referencial do nó 6 é delimitado pelos nós 1 e 3 ou seja, a veia de 6 é composta pelos nós 1, 3 e 6. Portanto, para que se entenda a anáfora expressa em [6] é preciso recuperar os segmentos de discurso correspondentes aos nós 1 e 3 (reproduzidos a seguir).

> [1] **A empresa Produtos Pirata Indústria e Comércio Ltda.**, de Contagem [3] deverá registrar este ano um crescimento de produtividade nas suas áreas comercial e industrial de 11% e 17%, respectivamente.

Em outras palavras, esse exemplo ilustra a construção das veias de cada *EDU* da árvore RST, a qual, por sua vez, indica as *EDUs* inter-relacionadas semanticamente. Se a estrutura RST do texto for coerente, as *EDUs* de uma mesma veia indicam seu elo semântico. Caso contrário, haverá *EDUs* semanticamente relacionadas em diferentes veias.

A *Veins Theory* pode ser usada tanto para guiar o processo de construção de estruturas RST (e, portanto, para interpretação textual), quanto para verificar se as estruturas RST resultantes da interpretação de um texto são, de fato, coerentes. Seretan e Cristea (2002), por exemplo, usam a *Veins Theory* para corrigir estruturas RST, de modo a assegurar que antecedentes e seus termos anafóricos apareçam no mesmo domínio de acessibilidade referencial. Sua proposta pode ser usada, por exemplo, em um ambiente automático de pós-edição de estruturas RST, o qual seria aplicável também a um ambiente de sumarização automática. Cristea et al. (2003) também sugerem que a *Veins Theory* pode mesmo restringir interpretações variadas de um mesmo discurso durante o processo de construção da sua estrutura RST.

Como já dito, neste trabalho a *Veins Theory* é usada para guiar o processo de poda de informações irrelevantes em uma estrutura RST de modo a assegurar que as estruturas RST dos sumários (isto é, as estruturas novas reconstruídas após a exclusão daquelas informações) não contenham quebras de co-referências, como será apresentado a seguir.

## 4. Heurísticas de Poda de Estruturas RST

No contexto de sumarização automática proposto, a poda de estruturas RST consiste em a) identificar informações irrelevantes ou menos salientes para exclusão da estrutura do sumário e b) verificar o relacionamento entre termos anafóricos e seus antecedentes, de modo a garantir a preservação do(s) antecedente(s) de um termo anafórico quando o mesmo for incluído na estrutura do sumário. Assim, as heurísticas de poda devem contemplar esses dois aspectos de natureza distinta para a sumarização automática, sendo o primeiro relacionado à informatividade e o segundo, à coerência do sumário.

Dessa maneira, as heurísticas se baseiam em duas hipóteses principais: a) a de que os satélites das relações podem ser supérfluos e, portanto, excluídos de uma estrutura RST de um sumário e b) a de que os satélites que contém os antecedentes dos termos anafóricos já inclusos na estrutura do sumário não podem ser excluídos.

A subseção a seguir apresenta a metodologia adotada para a especificação do elenco de heurísticas de poda para, então, apresentar o elenco de heurísticas propriamente dito, na subseção 4.2.

### 4.1 Metodologia Baseada em Corpus

A especificação das heurísticas de poda de estruturas RST baseou-se na análise de corpus, com dois propósitos: a) identificar construções discursivas cujos satélites apresentassem informações supérfluas e b) verificar os contextos co-referenciais que poderiam introduzir quebras de co-referência na sumarização automática. O corpus utilizado para análise é composto de 30 textos do gênero jornalístico, os quais foram extraídos do corpus TeMário (Pardo e Rino, 2003), disponível em http://www.linguateca.pt/Repositorio/TeMario. Cada texto do corpus foi analisado retoricamente com o auxílio da ferramenta *RST Annotation Tool*[4], resultando no corpus de suas estruturas RST (para mais detalhes sobre essa análise vide (Seno e Rino, 2004)).

Cada heurística tem o propósito, assim, de reestruturar estruturas RST de textos-fonte, considerando as restrições fundamentais de exclusão de informação irrelevante, ao mesmo tempo em que buscam a preservação da coerência. Logo, pode-se dizer que as heurísticas de sumarização de estruturas RST têm duas naturezas não excludentes: a informativa e a textual. A primeira, remetendo à identificação e exclusão de informações irrelevantes e a segunda, à garantia de textualidade das estruturas resultantes. Assim, a descrição metodológica a seguir distingue esses dois aspectos do trabalho.

#### 4.1.1 Análise com Foco na Informatividade

Primeiramente, compararam-se as estruturas RST de cada texto do corpus com seus correspondentes sumários manuais (SMs), os quais foram construídos por um profissional humano[5]. Essa comparação consistiu na verificação uma a uma das *EDUs* de uma estrutura RST de um texto que também estavam presentes no sumário manual. A hipótese, aqui, é que heurísticas baseadas na reprodução das informações constantes nos SMs garantam a informatividade mínima dos sumários automáticos, uma vez que os SMs são considerados ideais[6].

Verificaram-se, assim, as *EDUs* comuns a uma estrutura RST de um texto-fonte e ao seu sumário manual, além de se verificar o seu contexto. Esta verificação é necessária porque o inter-relacionamento retórico das *EDUs* a preservar no sumário automático também deve ser preservado, para que a mensagem subjacente permaneça inalterada (Rino, 1996). Isto poderia ser verificado, por exemplo, registrando-se as relações retóricas estabelecidas tanto na estrutura RST do texto-fonte quanto no sumário manual, assim como as informações satélites incluídas nos SMs. No entanto, neste caso, faz-se necessário também a construção das estruturas RST dos SMs. O levantamento dos satélites preservados nos SMs, assim como das relações retóricas envolvendo-os, é sumarizado na Tabela 3. A quarta coluna (Freqüência) indica a representatividade do satélite da relação em foco, com base nos SMs.

[4] Disponível em: http://www.isi.edu/~marcu/discourse/AnnotationSoftware.html.
[5] Esses sumários também compõem o TeMário.
[6] Para essa terminologia, vide (Mani, 2001).

Tabela 3: Representatividade dos satélites preservados nos SMs

| Relação Retórica | Ocorrência no Corpus | Satélites Preservados no SMs | Freqüência (%) |
|---|---|---|---|
| EXPLANTION ARGUMENTATIVE | 7 | 4 | 57 |
| MEANS | 2 | 1 | 50 |
| CAUSE | 12 | 6 | 50 |
| CONCESSION | 6 | 3 | 50 |
| EXPLANTION ARGUMENTATIVE-e | 6 | 3 | 50 |
| TEMPORAL AFTER | 4 | 2 | 50 |
| EXAMPLE | 4 | 2 | 50 |
| INTERPRETATION | 2 | 1 | 50 |
| JUSTIFY-e | 2 | 1 | 50 |
| RESULT | 4 | 2 | 50 |
| ELABORATION-e | 140 | 49 | 35 |
| COMPARISON | 9 | 3 | 33 |
| MEANS-e | 3 | 1 | 33 |
| REASON | 45 | 14 | 31 |
| EVIDENCE | 104 | 32 | 31 |
| ELABORATION | 413 | 119 | 29 |
| PURPOSE | 17 | 5 | 29 |
| CONDITION | 13 | 3 | 23 |
| JUSTIFY | 14 | 3 | 21 |
| ATTRIBUTION | 113 | 21 | 19 |
| PURPOSE-e | 6 | 1 | 17 |
| REASON-e | 6 | 1 | 17 |
| CIRCUMSTANCE-e | 8 | 1 | 13 |
| **PARENTHETICAL** | **13** | **0** | **0** |
| **COMPARISON-e** | **4** | **0** | **0** |
| **CIRCUMSTANCE** | **3** | **0** | **0** |
| **CAUSE-e** | **1** | **0** | **0** |
| **TEMPORAL SAME TIME** | **1** | **0** | **0** |
| **SUMMARY** | **1** | **0** | **0** |
| **SUMMARY-e** | **1** | **0** | **0** |

Como se pode notar, as relações retóricas ressaltadas em negrito na tabela não tiveram seus satélites preservados. Isto pode indicar que eles sejam irrelevantes para a sumarização e, portanto, a ocorrência de qualquer uma dessas relações pode indicar diretamente a exclusão de seu satélite das estruturas RST dos sumários. Outras relações como, por exemplo, *ELABORATION-e, COMPARISON, REASON,* que tiveram freqüência abaixo de 50%, também são significativas para as heurísticas de poda. As relações com representatividade de 50% ou mais poderiam levar a satélites que devem ser preservados nos sumários automáticos. No entanto, como a representatividade média dessas relações não é superior a 50%, elas também são consideradas na definição das heurísticas de sumarização.

Além das relações incluídas na Tabela 3, relações multinucleares também ocorrem no corpus. Entretanto, para a sumarização, elas não são muito significativas,

pois se consideradas em um sumário, todas as proposições inter-relacionadas devem ser incluídas, por apresentarem igual significância. Por essa razão, não há heurísticas para essas relações.

Devido ao tamanho limitado do corpus, buscou-se na literatura outros trabalhos que corroboram os resultados dessa análise. Por exemplo, Rino and Scott (1994) apontam, em seu trabalho, que os satélites das relações *CAUSE, ELABORATION, EXAMPLE*, *JUSTIFY* e *RESULT* apresentam informações pouco relevantes e podem ser excluídos em um sumário. Já Marcu (1998), em seu experimento, verificou que sujeitos humanos consideram satélites das relações *CIRCUMSTANCE, CONCESSION, CONDITION, EVIDENCE* e *EXAMPLE* irrelevantes para a sumarização.

No entanto, como este trabalho também é de natureza discursiva (vide seção 3), mais especificamente, visando evitar a quebra das cadeias de co-referências (CCRs), tarefas analíticas adicionais foram necessárias, como mostra a seção a seguir.

### 4.1.2 Análise com Foco na Textualidade

Visando, agora, a textualidade, isto é, a garantia de que as estruturas RST de sumários construídos automaticamente sejam coerentes e coesas (Rino, 1996), o corpus foi analisado especialmente com foco nas CCRs. Assim, buscou-se identificar como o domínio de acessibilidade referencial poderia contribuir para evitar a quebra de coerência já mencionada. Como visto anteriormente, esse domínio é delineado pelas veias de uma estrutura RST (seção 3).

Dessa forma, delimitaram-se as veias para cada uma das 30 estruturas RST dos textos do corpus. Além disso, suas CCRs foram anotadas usando a ferramenta MMAX (Müller and Strube, 2001)[7]. Após essa etapa de preparação dos dados, analisou-se para cada CCR de um texto se seu correspondente termo anafórico e antecedente estavam presentes em uma mesma veia. A hipótese, aqui, é que se uma CCR completa estiver presente em uma única veia, ao preservar toda a veia de uma *EDU,* quando a mesma for incluída em um sumário, não haverá quebra da CCR.

Com base nessa análise, observou-se que em 80% dos casos anáforas e antecedentes ocorrem em uma mesma veia. Assim, heurísticas baseadas na preservação das veias completas das *EDUs* inclusas na estrutura do sumário podem garantir a coerência mínima dos sumários automáticos.

Ambas as tarefas de análise do corpus permitiram a elaboração do elenco de heurísticas de poda, que será apresentado na seção a seguir.

## 4.2 Elenco de Heurísticas

O elenco de heurísticas de poda é composto por 30 heurísticas que visam identificar *EDUs* supérfluas em uma estrutura RST de um texto e excluir somente aquelas que não interferem na coerência, isto é, aquelas que estão fora do domínio de acessibilidade referencial de outras *EDUs*. De um modo geral, cada heurística se baseia na verificação da acessibilidade referencial de cada *EDU* já inclusa na estrutura do sumário. Portanto, todas elas terão, como ação, a exclusão de um satélite de uma relação retórica desde que este não esteja no domínio de acessibilidade referencial, isto é, na veia de uma *EDU* já inclusa. As heurísticas de poda são descritas a seguir, juntamente com sua descrição funcional e um exemplo (os satélites são apresentados em negrito)[8].

[7] Vide (Seno, 2004), para mais detalhes.
[8] Todos os exemplos foram extraídos do corpus escolhido.

Para simplificação, supõe-se que em todos os exemplos os satélites não pertençam às veias de outras *EDUs.*

H1 - Exclua y de attribution(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de attribution se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: A parceria de segurança é fundamental para manter a paz no Pacífico, especialmente nessa época de profundas mudanças na região, **disse o presidente americano, durante uma entrevista à imprensa concedida ao lado do primeiro-ministro japonês, Ryutaro Hashimoto.**

H2 - Exclua y de cause(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de cause se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Nos EUA, há cerca de 200 milhões de armas. **O índice de assalto nos EUA é cerca de 130 vezes superior ao do Japão.**

H3 - Exclua y de cause-e(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de cause-e se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Devido a mudanças legais, facilitaram-se as aposentadorias de professores, **o que fez aumentar a despesa do ministério com inativos.**

H4 - Exclua y de circumstance(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de circumstance se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: O conflito começou pouco depois das 16h, **quando 150 policiais militares chegaram à área onde estavam acampados cerca de 1.500 sem-terra.**

H5 - Exclua y de circumstance-e(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de circumstance-e se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Diz o pesquisador de religiões Joaquim de Andrade, 32, que alguns membros dessas seitas têm sustentado que, **quando finalmente o apocalipse chegar**, só haverá vagas no céu para um número justo de 144 mil pessoas puras.

H6 - Exclua y de comparison(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de comparison se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Na França, a média de fecundidade é de 1,3 filho por mulher. **Para efeito de comparação, em São Paulo, segundo a demógrafa Bernadete Waldvogel, do Seade, a média é de 2,2 filhos por mulher.**

H7 - Exclua y de comparison-e(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de comparison-e se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Como era de se esperar, a municipalização encontrou e encontra muitas resistências - **comparáveis às encontradas no programa de privatização.**

H8 - Exclua y de concession(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário

Função: Excluir satélite de concession se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: O príncipe promete não se casar de novo, **apesar de ter sido visto em público, no ano passado, em companhia de sua amante, Camilla Parker-Bowles.**

H9 - Exclua y de condition(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de condition se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Um editorial no jornal do Partido Comunista da Coréia do Norte, Rodong Sinmun dizia ontem que os vizinhos do sul enfrentariam um desastre irrevogável **caso ignorassem os alertas de Pionguiangue sobre o que considerava ser movimentações beligerantes.**

H10 - Exclua y de elaboration(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de elaboration se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Descendente de africâners (colonizadores de origem holandesa), De Klerk nasceu em Johannesburgo, em 18 de março de 1936. **Seu pai foi membro do Partido Nacional (PN), fundado em 1948.**

H11 - Exclua y de elaboration-e(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de elaboration-e se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Os sul-coreanos tentaram ontem angariar o apoio mundial à sua denúncia contra a Coréia do Norte pela violação da trégua acertada em 1953, **que acabou com a guerra entre os dois países.**

H12 - Exclua y de evidence(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de evidence se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Em abril do ano passado, num gesto surpreendente, De Klerk disse adeus a seu passado racista e pediu desculpas pela apartheid que vigorou de 1948 a 1989. **"Não era nossa intenção privar as pessoas de seus direitos e causar miséria, mas a segregação e a apartheid levaram exatamente a isso e eu lamento profundamente, afirmou".**

H13 - Exclua y de example(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de example se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Uma menor fertilidade pode trazer impactos consideráveis na qualidade de vida **como, por exemplo, provendo mais educação, atendimento de saúde e oportunidades de empregos.**

H14 - Exclua y de explanation_argumentative(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de explanation_argumentative se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Até o fim do século o mundo vai assistir ao fenômeno da desmetropolização, **ou seja, a tendência desta década será a desconcentração populacional das metrópoles.**

H15 - Exclua y de explanation_argumentative-e(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de explanation_argumentative-e se não estiver na veia de uma *EDU* já selecionada para o sumário

Exemplo: Em outubro cai um pilar do apartheid - **a lei que dividia locais públicos entre brancos e negros.**

H16 - Exclua y de interpretation(x,y) se y ∉ veia de z, para alguma *EDU* z ∈ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de interpretation se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: No caso específico do Brasil, a expectativa dos cientistas é que a partir de 2020 o país vá ter seu crescimento populacional estabilizado e, por volta de 2050, essa taxa chegará a zero. **Isso significa que o número de mortes vai se igualar ao de nascimentos.**

H17 - Exclua y de justify(x,y) se y ∉ veia de z, para alguma *EDU* z ∈ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de justify se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Dessa maneira, analfabetos são mais propensos a fumar, consumir álcool, viver de maneira sedentária e serem obesos. **A baixa instrução implica menor esclarecimento médico e menos cuidado com a saúde.**

H18 - Exclua y de justify-e(x,y) se y ∉ veia de z, para alguma *EDU* z ∈ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de justify-e se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: De acordo com Will, as pessoas devem ser sempre lembradas de que a passagem mais barata envolve algum risco **- os preços reduzidos são geralmente uma decorrência de cortes nas revisões dos aparelhos e redução do tempo de treinamento das tripulações.**

H19 - Exclua y de means(x,y) se y ∉ veia de z, para alguma *EDU* z ∈ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de means se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Ao completar sessenta anos de fundação, no mesmo dia do aniversário da cidade, a universidade responsável por quase metade dos doutoramentos do país pretende ampliar mesmo é sua participação nos grandes debates nacionais. **Através do Instituto de Estudos Avançados (IEA), a USP pretende discutir e apresentar propostas para questões como a Amazônia e o sistema Judiciário do país.**

H20 - Exclua y de means-e(x,y) se y ∉ veia de z, para alguma *EDU* z ∈ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de means-e se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: A idéia do relator da revisão constitucional era tão-somente flexibilizar os monopólios, **através de concessão de serviços.**

H21 - Exclua y de parenthetical(x,y) se y ∉ veia de z, para alguma *EDU* z ∈ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de parenthetical se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: A pesquisa sobre a situação educacional no mundo está incluída num relatório intitulado The Progress of Nations **(O Progresso das Nações).**

H22 - Exclua y de purpose(x,y) se y ∉ veia de z, para alguma *EDU* z ∈ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de purpose se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Em novembro de 92, o presidente propõe um amplo programa de negociações **para a realização das primeiras eleições multirraciais da África do Sul.**

H23 - Exclua y de purpose-e(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de purpose-e se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Os dois só evitam conversar sobre guerrilha do Araguaia, **para evitar constrangimentos.**

H24 - Exclua y de reason(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de reason se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Não por acaso, De Klerk e Mandela, ex-inimigos, dividiram o prêmio Nobel da Paz de 1993. **O ex-racista De Klerk libertou o ex-extremista Mandela em 1990, após 27 anos de prisão.**

H25 - Exclua y de reason-e(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de reason-e se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Mario Silva Ramos acredita no fim do mundo para o ano de 1999 **- graças a uma frase que ele retirou do livro do Apocalipse: um raio branco varrerá os não convertidos do centro da terra e só serão arrebatados ao Paraíso os merecedores.**

H26 - Exclua y de result(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de result se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Com a velocidade do mercado, inúmeros produtos, serviços e negócios simplesmente desapareceram **pela incapacidade de percepção ou adaptação às novas expectativas de consumo.**

H27 – Exclua y de summary(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de summary se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Apesar do consenso sobre a necessidade de reformular a Previdência Social, as propostas hoje em discussão apresentam pontos divergentes... **Para viabilizar as mudanças na Previdência, serão necessárias mudanças na Constituição aprovada em 1988.**

H28 – Exclua y de summary-e(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de summary-e se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: O atendimento médico, psicológico ou mesmo odontológico na USP nas unidades de ensino é mais dirigido às necessidades de ensino e pesquisa **- em suma, trata-se de uma troca entre a população, estudantes e pesquisadores.**

H29 – Exclua y de temporal_after(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário
Função: Excluir satélite de temporal_after se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Foi a primeira ação da Otan contra sérvios **desde o ataque aéreo às suas posições no encrave de Gorazde, em abril.**

H30 – Exclua y de temporal_same_time(x,y) se y $\notin$ veia de z, para alguma *EDU* z $\in$ conjunto de *EDUs* já selecionadas para o sumário

Função: Excluir satélite de temporal_same_time se não estiver na veia de uma *EDU* já selecionada para o sumário
Exemplo: Para Annateresa, esse discurso impediu a avaliação de características importantes nas obras de Anita, Tarsila, Di Cavalcanti (1897-1976), Vicente do Rêgo Monteiro (1899-1970), Lasar Segall (1891-1957) e Oswaldo Goeldi (1895-1961). **Ao mesmo tempo, consagrou pintores apenas por mérito temático.**

A seção a seguir descreve o sumarizador automático de estruturas RST baseado nessas heurísticas.

## 5. O Sumarizador Automático Baseado em Heurísticas de Poda de Estruturas RST

A Figura 6 ilustra a arquitetura de um protótipo de sumarizador de estruturas RST composto de três módulos de processamento em *pipeline*. Neste trabalho, o foco principal está na delimitação das veias e na poda das estruturas RST dos textos (e, portanto, nos dois primeiros módulos). Supõe-se que uma ou mais heurísticas de poda possam ser aplicadas em uma estrutura RST de um texto-fonte, resultando na estrutura RST do seu sumário. O processo de realização lingüística proposto é elementar, baseado em *templates*, sendo vários deles incorporados de trabalho anterior (Pardo, 2002). Claramente, a construção da estrutura superficial baseada em *templates* para a expressão lingüística das estruturas RST dos sumários não é uma medida adequada, podendo levar a resultados insatisfatórios. Porém, o foco do trabalho atual não está na realização lingüística, mas na estruturação discursiva segundo as duas naturezas previstas (vide seção 4). Certamente, este fator alterará as possíveis formas de avaliação da proposta, questão a ser discutida no futuro.

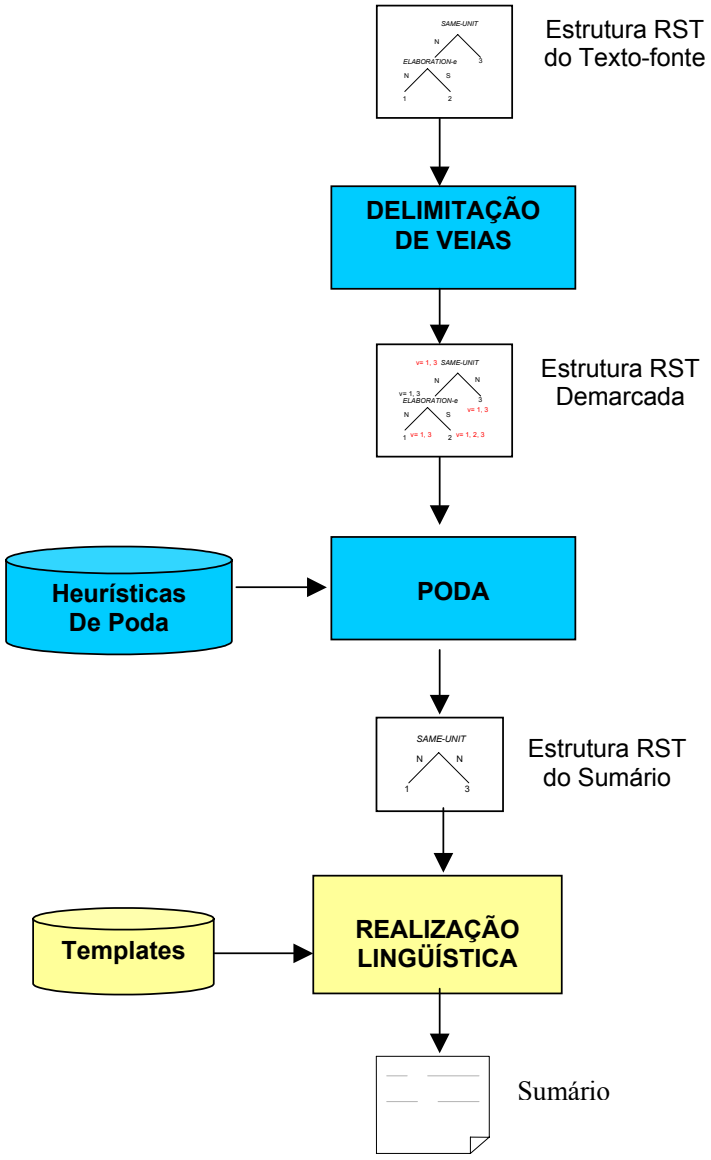


Figura 6: Arquitetura do sumarizador automático de estruturas RST

A seguir, apresenta-se o principal processo do sumarizador automático: o mecanismo de poda.

### 5.1 Mecanismo de Poda

O processo de poda do sumarizador de estruturas RST é composto por três passos principais: a) classificação das *EDUs* com base na nuclearidade e na profundidade em que se encontram na estrutura RST; b) aplicação das heurísticas de poda e c) seleção das *EDUs* para o sumário com base em taxa de compressão previamente estabelecida. As subseções 5.1.1, 5.1.2 e 5.1.3 descrevem cada um desses passos, respectivamente, para, então, apresentar um exemplo de poda, na subseção 5.3.

### 5.1.1 Classificação de EDUs

O processo de poda classifica todas as *EDUs* da estrutura RST com o propósito de obter uma classificação de importância dessas *EDUs* e, assim, estabelecer uma prioridade para a aplicação das heurísticas de poda. O algoritmo usado para classificação foi proposto por Marcu (1997, 1999) e parte do princípio de que as *EDUs* que estão mais próximas da raiz da árvore RST são mais importantes (ou salientes) do que aquelas que estão em níveis mais profundos.

Assim, o algoritmo atribui a cada nó interno da árvore RST um conjunto promocional (*promotion set*) formado por suas unidades salientes (ou promocionais), isto é, as unidades mais importantes no segmento de discurso correspondente. O conjunto promocional equivale aos *heads* da *Veins Theory*, conforme visto na seção 3, e é determinado de maneira *bottom-up*, como segue:

- A unidade mais saliente de um nó folha é o próprio nó folha;
- As unidades mais salientes de um nó interno são formadas pela união das unidades mais salientes dos filhos nucleares imediatos do referido nó;

Dessa forma, o cômputo da saliência das *EDUs* de uma árvore RST se baseia tanto na nuclearidade quanto em sua profundidade na árvore. Além disso, as *EDUs* mais próximas da raiz têm um *score* maior do que outras *EDUs* mais profundas. Aplicando-se repetidamente esse método de cálculo da saliência a cada nó de uma árvore RST, obtém-se a ordem de importância de todas as suas *EDUs*.

O *score* de importância *s(u,D,d)* de uma unidade *u* em uma estrutura de discurso *D,* a uma profundidade *d,* pode ser definido, pela seguinte função recursiva:

$$s(u,D,d) = \begin{cases} 0 & \text{se } D \text{ é nula,} \\ d & \text{se } u \in \text{prom(D),} \\ d\text{-}1 & \text{se } u \in \text{paren(D),} \\ max(s(u, C(D), d\text{-}1)) & \text{caso contrário.} \end{cases}$$

para:

prom(D) é o conjunto promocional do nó D
paren(D) é o conjunto de unidades pais do nó D
C(D) é o conjunto de subárvores filhas do nó D

### 5.1.2 Aplicação das Heurísticas de Poda

Após classificar as *EDUs* com base na função de saliência, as heurísticas de poda podem ser aplicadas à estrutura RST, para sua sumarização. Assim, o mecanismo de poda verifica para cada uma das *EDUs* da estrutura, segundo sua ordem de importância, se há alguma heurística que seja aplicável. Caso haja alguma, então, esta é aplicada e a relação retórica envolvendo-a é excluída, para a reestruturação da árvore RST do sumário.

Ao fim da poda da estrutura RST, obtém-se uma estrutura RST do sumário. No entanto, para o sumarizador proposto (vide Figura 6), essa estrutura resultante não é a estrutura final, pois supõem-se, ainda, que seja possível estipular a taxa de compressão, como é usual na maioria dos casos de sumarização automática. Neste caso, uma segunda etapa de estruturação tem lugar, como segue.

### 5.1.3 Seleção de EDUs

Embora se trate de um sumarizador fundamental, isto é, que processa, sobretudo, a estrutura discursiva de um sumário, considera-se, aqui, que a taxa de compressão poderá ser usada também no nível profundo, para expressar o volume aproximado de unidades informativas que o sumário suposto irá conter. É importante observar que, nesse nível, somente é possível calcular o número de unidades informativas (e não de palavras), sendo que o tamanho real do sumário final não pode ser delineado pelo módulo de poda, mas somente pelo realizador lingüístico, ou seja, na última etapa de um sumarizador automático completo.

Assim, depois de aplicar as heurísticas de poda, a taxa de compressão servirá para determinar o número de *EDUs* que comporão a estrutura RST sumarizada. A restrição fundamental, nesse passo, é que a seleção de uma *EDU* qualquer implica a seleção de todas as *EDUs* que compõem a sua veia, mesmo que, em alguns casos, isso provoque um maior número de *EDUs* na estrutura resultante. É válido dizer que, no protótipo proposto, a taxa de compressão é estabelecida pelo próprio usuário.

Para melhor ilustrar esse processo, a seção a seguir apresenta um exemplo de poda de uma estrutura RST.

## 5.2 Exemplo de Poda

Considere a árvore RST do texto-exemplo da Figura 4 (seção 3) reproduzida na Figura 7, por conveniência. O primeiro passo do processo de poda é a classificação das *EDUs* com base na função de saliência. A(s) *EDU*(s) mais saliente(s) de um dado segmento é apresentada em negrito na figura. De acordo com o algoritmo de Marcu, a ordem de importância das *EDUs* é a seguinte: 1, 3> 8 > 6, 9, 10> 4, 7> 5> 2. Nesta representação, cadeias de *EDUs* separadas por vírgula indicam que elas têm o mesmo grau de importância.

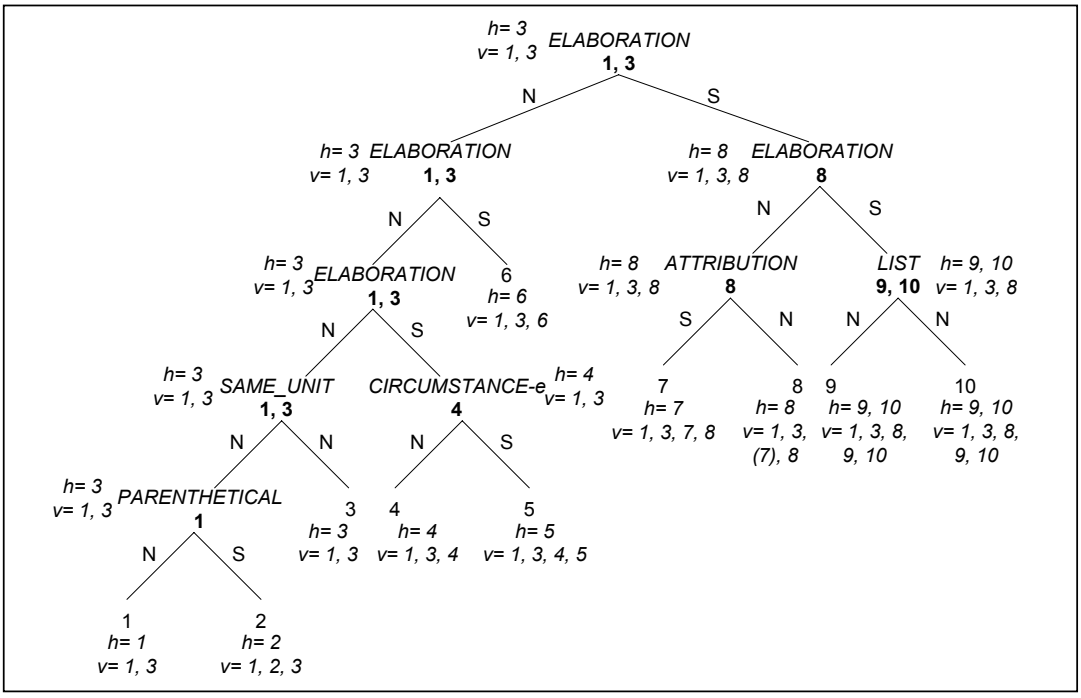


Figura 7: Classificação das EDUs da árvore RST do texto-exemplo (vide Figura 4)

Após classificar as *EDUs*, o segundo passo é verificar para cada *EDU* se há alguma heurística aplicável. Neste caso, apenas para as *EDUs* 6, 5 e 2 há heurísticas que se aplicam, são elas: H10, H5 e H21, respectivamente, as quais remetem às relações retóricas *ELABORATION, CIRCUMSTANCE-e* e *PARENTHETICAL*. Ao aplicar as heurísticas essas relações também são excluídas e a árvore podada é reestruturada. Assim, tem-se como resultado a estrutura RST ilustrada na Figura 8, na qual a ordem de importância das *EDUs* é preservada.

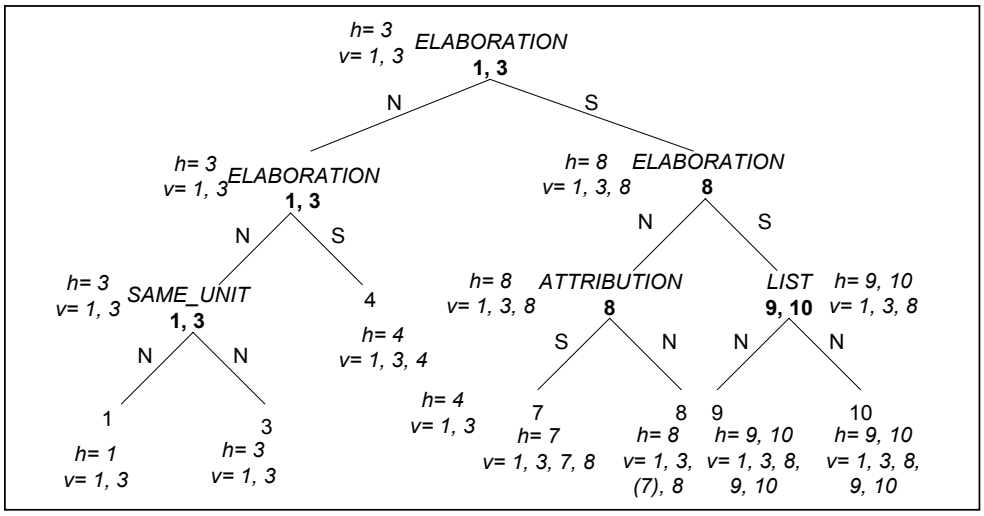


Figura 8: Aplicação das heurísticas de poda na árvore RST do texto-exemplo

Como último passo do processo de poda, essa estrutura é adequada à taxa de compressão. Considerando-se uma taxa de compressão de 70% (ou seja, uma estrutura RST de um sumário composta de 30% das *EDUs* do texto-fonte), a estrutura RST do sumário pretendido resulta na ilustrada na Figura 9. Vale notar que a ordem de importância das *EDUs* e a restrição de se incluir uma veia completa, quando uma de suas *EDUs* for incluída devem ser mantidas. Como o texto-exemplo é composto por 10 *EDUs*, aplicando-se a taxa de compressão, a estrutura final teria somente três *EDUs*: 1, 3 e 8. Porém, como a *EDU* 7 está na veia da *EDU* 8 esta também é selecionada, resultando em uma estrutura RST com quatro *EDUs*: 1, 3, 7 e 8.

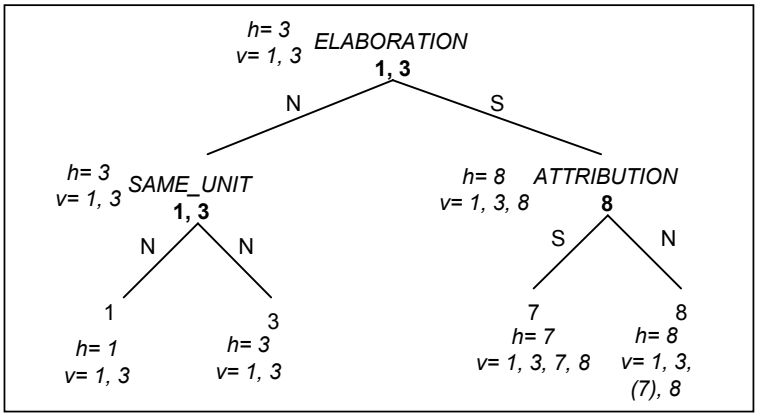


Figura 9: Estrutura RST do sumário

Uma possível realização lingüística para esta estrutura seria a apresentada na Figura 10. Esse processo seria similar ao processo de se aplicar *templates* que indicam como realizar as relações retóricas, "congelando-se" as *EDUs* a suas unidades frasais, como aparecem no texto-fonte. A elaboração de um módulo real de realização lingüística não é considerada neste trabalho, como já foi explicitado na seção 5.

A empresa Produtos Pirata Indústria e Comércio Ltda. de Contagem deverá registrar este ano um crescimento de produtividade nas suas áreas comercial e industrial de 11% e 17%, respectivamente.
A coordenadora d**o programa de qualidade** na empresa, Márcia Cristina de Oliveira Neto, disse que ainda não é possível dimensionar os ganhos financeiros que "certamente" a empresa terá, em conseqüência da melhoria da qualidade de seus produtos e serviços.

Figura 10: Sumário do texto-exemplo

Vale notar que o sumário apresentado na Figura 10 apresenta uma quebra de co-referência (em negrito na figura). Isso ocorre porque não há uma relação estrutural entre a *EDU* que contém o termo anafórico "**o programa de qualidade**", *EDU* 7, e a *EDU* que contém o seu antecedente "**o Programa Sebrae de Qualidade Total**", *EDU* 5 (conforme ilustrado na Figura 7), caso esse em que a *Veins Theory* não prevê tratamento (Cristea et. al, 1998). Apesar dessa quebra de co-referência não prevista pela *Veins Theory* e dessa análise preliminar, as heurísticas apontam resultados promissores, uma vez que a mensagem principal do texto-exemplo (Figura 4) e a coerência foram preservadas.

## 6. Considerações Finais

Este relatório apresentou um elenco de heurísticas de poda para a sumarização automática de estruturas RST de textos. Além do conjunto de heurísticas, também foram apresentadas a metodologia de desenvolvimento e a metodologia de aplicação dessas heurísticas no protótipo do sumarizador automático de estruturas RST.

O elenco de heurísticas de poda é composto por 30 heurísticas responsáveis por identificar informações irrelevantes na estrutura RST de um texto e excluir somente aquelas que não prejudicam a coerência, em relação a cadeias de co-referências (CCRs).

É válido dizer que em alguns casos as heurísticas de poda não são suficientes para garantir a preservação de todas as CCRs, pois há casos onde a *EDU* que contém o antecedente de um termo anafórico não tem relação estrutural com a *EDU* que contém a anáfora, casos esses que a própria *Veins Theory* não fornece tratamento.

## Referências Bibliográficas

Cristea, D.; Ide, N.; Romary, L. (1998). Veins Theory: A Model of Global Discourse Cohesion and Coherence. *In the Proceedings of the Coling/ACL' 1998*, pp.281-285. Montreal, Canadá.

Cristea, D.; Postolache, O.; Puscasu, G.; Ghetu, L. (2003). Summarizing Documents Based on Cue-phrases and References. *In the Proceedings of the International Symposium on Reference Resolution and its Apllications to Questions Answering and Summarization.* Veneza.

Grosz, B.; Joshi, A.; Weinstein, S.; (1995). Centering: a Framework for Modelling the Local Coherence of Discourse. *Computational Linguistic 21 (2)*, pp. 203-225, June.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization.* Technical Report ISI/RS-87-190.

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.* PhD Thesis, Department of Computer Science, University of Toronto.

Marcu, D. (1998). To build text summaries of high quality, nuclearity is not sufficient*. The Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization,* pages 1-8, Stanford, CA.

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

Marcu, D (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. Cambridge, Massachusetts.

Müller C. and Strube, M. (2001). MMAX: A tool for the annotation of multi-modal corpora. *In the Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*. Aalborg, Denmark, pp. 90-95.

O'Donnell, M. (1997). Variable-Length On-Line Document Generation. *In the Proceedings of the 6th European Workshop on Natural Language Generation*, Gerhard-Mercator University, Duiburg, Germany.

Ono, K.; Sumita, K.; Miike, S. (1994). Abstract Generation Based on Rhetorical Structure Extraction. *In the Proceedings of the International Conference on Computational Linguistic – Coling-94*, pp 344-348, Japan.

Paraboni, I. (1997). *Uma arquitetura para a Resolução de Referências Pronominais Possessivas no Processamento de Textos em Língua Portuguesa*. Dissertação de Mestrado. PUCRS, Porto Alegre-RS.

Pardo, T. A. S. (2002). *DMSumm: Um gerador automático de sumários*. Dissertação de Mestrado. UFSCar, São Carlos-SP.

Pardo, T.A.S. e Rino, L.H.M. (2003). *TeMário: Um Corpus para a Sumarização Automática de Textos*. Série de Relatórios Técnicos: NILC-TR-03-09, ICMC/USP, São Carlos-SP.

Rino, L.H.M. and Scott, D. (1994). *Automatic Generation of Draft Summaries: Heuristics for Content Selection*. ITRI-94-8 Technical Report. University of Brighton, UK.

Rino, L.H. M. (1996). *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. Tese de Doutorado. IFSC-USP São Carlos – SP.

Sparck Jones, K. (1993). *Discourse Modelling for Automatic Summarising*. Tech. Rep. No. 290. University of Cambridge, February.

Seno, E.R.M. (2004). *Especificação de Heurísticas de Sumarização de Estruturas RST com base na Preservação dos elos Co-referenciais*. Monografia de Qualificação de Mestrado. UFSCar, São Carlos-SP.

Seno, E.R.M. e Rino, L.H.M. (2004). *Análise Discursiva para a Sumarização Automática de Textos em Português*. Série de Relatórios Técnicos: NILC-TR-04-06, ICMC/USP, São Carlos-SP.

Seretan, V.; Cristea, D. (2002). The Use of Referential Constraints in Structuring Discourse. *In the Proceedings of the LREC'2002*. Las Palmas, Spain.

Vieira, R.; Salmon-Alt, S.; Schang, E. (2002). Multilingual Corpora Annotation for Processing Definite Descriptions. *In the Proceedings of the Portugal for Natural Language Processing – PorTAL – 2002*, Faro, Portugal.