UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM

CIÊNCIA DA COMPUTAÇÃO

"RHeSumaRST: Um Sumarizador Automático de Estruturas RST"

Eloize Rossi Marques Seno

São Carlos
Agosto/2005

AGRADECIMENTOS

A Deus, por ter me dado força e coragem para enfrentar os momentos mais difíceis.

A meu esposo Wesley, pelo companheirismo, paciência e compreensão.

A minha família, que sempre me apoiou.

A minha orientadora, pelas orientações tanto profissionais como pessoais, pelas contribuições com seu conhecimento e experiência, pelo profissionalismo, por sua dedicação sem medir esforços e pela amizade.

Ao Leandro M. Hanada, pela contribuição tão importante para este trabalho.

Aos colegas do NILC e do LIAA, pelas contribuições, pelos seminários muito produtivos, pelos bate- papos que nada contribuíram para este trabalho, pelos cafezinhos e pelas amizades.

Ao NILC – Núcleo Interinstitucional de Lingüística Computacional, pelo apoio e uso de suas instalações.

INDICE GERAL

INDICE DE FIGURAS

INDICE DE TABELAS

RESUMO

Este trabalho apresenta um modelo de sumarização automática que se baseia no modelo de estruturação de discurso *Rhetorical Structure Theory – RST* e no modelo de coerência global do discurso *Veins Theory – VT*. A RST permite a estruturação de um discurso relacionando-se unidades discursivas com base em relações retóricas, isto é, permitindo recuperar as relações de significados entre tais unidades. Com base na estruturação RST, a *Veins Theory* delimita o domínio de acessibilidade referencial para cada unidade do discurso na forma de "veias", indicando os limites nos quais os antecedentes de uma anáfora podem ocorrer ao longo do discurso. Além dessas teorias, o modelo também incorpora o modelo de classificação de saliência de unidades discursivas proposto por Marcu (1997a), que obtém uma ordem de importância das unidades discursivas de uma estrutura RST.

O modelo de sumarização proposto consiste em um elenco de heurísticas que visam identificar informações supérfluas em uma estrutura RST de um texto, para exclusão durante a construção do seu sumário, tendo sempre como foco a preservação dos elos co-referenciais. Dessa forma, as heurísticas são guiadas por características específicas das relações retóricas da RST e por restrições de acessibilidade referencial da *Veins Theory*. Assim, o sumarizador proposto se resume à poda de segmentos discursivos irrelevantes das estruturas RST de textos, resultando em seus correspondentes sumários. As principais contribuições deste trabalho são a proposta de um novo modelo de sumarização automática e um protótipo para a sua aplicação automática.

ABSTRACT

This work presents an automatic summarization model based on both the *Rhetorical Structure Theory – RST* – and the *Veins Theory – VT*. RST allows inter-relating discourse units by means of rhetorical relations. These, in turn, mirror meaning relations between those units. Adding to RST, VT delimits the domain of referential accessibility of each discourse unit of an RST tree, resulting in its "vein". A vein signals, thus, the limits of a discourse unit that may enclose its anaphora antecedents. The automatic summarization model also embeds Marcu's model of salience: once a discourse is structured as an RST tree, its units are classified according to their salience by considering the its deep in the tree.

The model consists of a set of pruning heuristics that aim at identifying superfluous information in an RST tree of a text. In excluding them, the resulting summary RST tree and, thus, the text summary, should preserve the co-referential chains. In this way, the heuristics are driven by both, specific features of RST relations and constraints on the referential accessibility provided by VT. The main contributions of this work include the proposal of the AS model itself and the availability of a prototype for its automatic exploration.

1. INTRODUÇÃO

Com o crescente volume de informações disponíveis, principalmente em formato eletrônico, e o tempo cada vez menor que as pessoas dispõem para ler e absorver o máximo dessas informações, os sumarizadores automáticos de textos têm desempenhado um papel bastante importante. Os sumários podem ser muito úteis, por exemplo, quando o leitor deseja identificar os documentos mais relevantes para o seu propósito dentre um conjunto de documentos.

A sumarização automática de textos pode ser vista como um processo de identificação do conteúdo mais relevante em um texto e posterior condensação, preservando a mensagem original pretendida. No entanto, um dos grandes problemas da sumarização automática (SA) é identificar quais informações são mais importantes em um texto para compor o seu sumário. Um modelo que tem sido muito utilizado com esse objetivo é a *Rhetorical Structure Theory – RST* (Mann and Thompson, 1987) (por exemplo, Sparck Jones, (1993a); Ono et al., (1994); Rino, (1996); O'Donnel, (1997); Marcu (1997a, 1997b, 1999)).

A RST permite a estruturação de um texto relacionando suas unidades discursivas por meio de relações retóricas. Por hipótese, se o texto for coerente, sua estrutura RST (ou estrutura retórica) permitirá recuperar sua mensagem e, assim, poderá ser usada para identificar as informações supérfluas, para exclusão durante a construção de um sumário. Um dos trabalhos mais relevantes que contempla a RST para a SA é o de Marcu (1997a). Nesse trabalho é proposto um modelo de SA que se baseia nas estruturas retóricas de textos para computar a saliência de cada unidade discursiva, considerando que as unidades mais salientes apresentam informações mais relevantes do texto, sendo adequadas, portanto, para compor o sumário. A saliência das unidades discursivas é determinada com base em sua profundidade na árvore retórica (vide Capítulo 5).

Uma limitação da RST é que ela não propõe tratar fenômenos lingüísticos como, por exemplo, o relacionamento entre as cadeias de co-referências (CCRs). As CCRs são referências a uma entidade já introduzida na comunicação e reproduzidas ao longo do discurso (Milner, 2003). A forma mais comum de co-referência é a relação anafórica, estabelecida entre uma expressão de referência (o termo anafórico ou, simplesmente,

anáfora) e um termo que a antecede no discurso (o termo antecedente ou, simplesmente, antecedente).

A limitação da RST, em relação ao fenômeno de co-referenciação, é que, embora ela possa indicar informações supérfluas em geral correspondentes aos satélites na estrutura RST, ela não é capaz de indicar as informações cuja exclusão implique a quebra das CCRs. Assim, uma unidade discursiva que contém uma anáfora pode ser escolhida para compor o sumário, mas a unidade discursiva que contém o seu antecedente não. Neste caso, o sumário será incoerente, podendo ser incompreensível para o leitor. Por exemplo, se o texto "João estudou durante todo o final de semana. Ele foi muito bem na prova." for sumarizado, resultando somente na sentença "Ele foi muito bem na prova.", o sumário pode não ser inteligível, devido à inclusão da anáfora "ele" sem o seu antecedente "João".

Uma proposta de superar a limitação de se excluir informações somente com base na RST é apresentada pela *Veins Theory* (VT) (Cristea et al., 1998), que delimita 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 o antecedente de uma anáfora pode ocorrer ao longo do discurso. Dessa forma, a *Veins Theory* pode ser usada para determinar as unidades discursivas que podem levar a um sumário coerente.

Assim, este trabalho apresenta um modelo de SA que agrega o potencial de estruturação da RST à proposta de relacionamento co-referencial da *Veins Theory*, sob a hipótese de que a exclusão (ou poda) de unidades discursivas irrelevantes em uma estrutura RST pode ser guiada por restrições de acessibilidade referencial, garantindo, dessa forma, a produção de sumários mais coerentes. O modelo também adiciona a proposta de classificação de saliência de unidades discursivas do modelo de Marcu, como um auxílio na identificação de informações irrelevantes. O modelo de sumarização proposto consiste em um conjunto de heurísticas que focalizam as relações retóricas em uma estrutura RST, procurando identificar aquelas que representam as informações menos relevantes do discurso para exclusão, de modo a preservar a coerência dos sumários. Propõe-se a implementação do protótipo RHeSumaRST (sigla para Regras Heurísticas de Sumarização de estruturas RST) com base nesse modelo de sumarização profundo. Portanto, como será mostrado no decorrer deste trabalho, a proposta de sumarização é essencialmente baseada na modelagem lingüística e discursiva.

No próximo capítulo apresenta-se a motivação deste trabalho contextualizando-o com base na sumarização humana e na SA, segundo a abordagem profunda. Nos Capítulos 3 e 4 são descritas as teorias RST e *Veins Theory*, respectivamente. No Capítulo 5 apresenta-se o modelo proposto por Marcu, dentre outros trabalhos de SA que contemplam a RST ou a *Veins Theor*y e que, de alguma forma, relacionam-se com o modelo aqui proposto. No Capítulo 6 apresenta-se o RHeSumaRST, bem como a metodologia de desenvolvimento adotada para a sua especificação e o elenco de heurísticas de poda. Alguns experimentos realizados com o RHeSumaRST são apresentados no Capítulo 7. No Capítulo 8 apresentam-se as principais contribuições deste trabalho e algumas possibilidades de trabalhos futuros. Por fim, o Capítulo 9 apresenta as considerações finais do trabalho e suas principais limitações.

2. A SUMARIZAÇÃO AUTOMÁTICA BASEADA EM CONHECIMENTO PROFUNDO

2.1 Caracterização

A abordagem profunda de SA toma do processamento humano da linguagem (e, portanto, da psicologia cognitiva) a base interpretativa para a compreensão do texto-fonte e posterior condensação/produção do(s) sumário(s) correspondente(s) (Sparck Jones, 1993a, 1993b). Segundo essa abordagem, ela baseia-se em teorias lingüísticas e modelos formais para simular o processo humano de sumarização. No entanto, simular a sumarização humana é uma tarefa de grande complexidade, pois envolve processos igualmente complexos de interpretação do texto e representação somente das informações mais relevantes, relacionadas à sua idéia central (Rino e Pardo, 2003). Desta forma, é necessário que se estabeleçam fundamentos e estratégias claras sobre o processo de sumarização humana de modo a permitir a modelagem automática. Neste sentido, a próxima seção descreve como profissionais humanos procedem na tarefa de sumarização para, então, apresentar a arquitetura de um sistema de SA baseado na abordagem profunda, na seção 2.3.

2.2 Sumarização Humana

A sumarização humana pode ser definida como a tarefa de identificar o que é relevante em um texto e, então, traçar um novo texto preservando a sua idéia principal, sem perder o significado original pretendido (Rino e Pardo, 2003). É importante ressaltar a complexidade e diversidade características dessa tarefa: distinguir os graus de importância das informações depende de vários fatores: dos interesses do autor do sumário, para a transmissão de sua mensagem ao leitor, dos objetivos ou interesses de seus possíveis leitores e da importância relativa (e subjetiva) que o próprio autor (ou leitor) atribui às

informações textuais. Por exemplo, o texto-exemplo da Figura 1[1] (com as sentenças numeradas, para referência) pode ter as manchetes M1 e M2 (apresentadas na Figura 2), as quais ressaltam informações diferentes, atribuindo diferentes graus de relevância ao conteúdo original do mesmo texto-fonte. É válido notar que o foco da manchete M1 está explícito no texto, na sentença [1]. Porém, o foco da manchete M2 está implícito no texto, tendo sido inferido do conjunto de sentenças [1] e [2].

[1] A empresa Produtos Pirata Indústria e Comércio Ltda., de Contagem (na região metropolitana de Belo Horizonte), deverá registrar este ano um crescimento de produtividade nas suas áreas comercial e industrial de 11% e 17%, respectivamente. [2] Os ganhos são atribuídos pela diretoria da fábrica à nova filosofia que vem sendo implantada na empresa desde outubro do ano passado, quando a Pirata se iniciou no Programa Sebrae de Qualidade Total.
[3] Dona de 65% do mercado mineiro de temperos, condimentos e molhos, a Pirata reúne atualmente 220 funcionários. [4] A coordenadora do 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. [5] Por enquanto, os benefícios mais visíveis, segundo ela, são a organização e a limpeza da fábrica. [6] "Também a relação entre as pessoas tem melhorado bastante. [7] As informações estão mais claras e os funcionários e clientes, mais satisfeitos".

Figura 1: Texto-exemplo 1

M1: Empresa mineira deve registrar este ano um aumento de produtividade nas áreas comercial e industrial de 11% e 17%, respectivamente.

M2: Empresa mineira implanta Programa de Qualidade Total e obtém aumento significativo de produtividade.

Figura 2: Manchetes do texto-exemplo 1

Esse exemplo permite-nos identificar algumas possíveis características da sumarização humana: variações de conteúdo informativo, variações no tamanho (comprimento) dos sumários, diversidade de sumários para um mesmo texto-fonte e influência da intenção comunicativa do autor do sumário. Além dessas características, pode-se citar ainda a variação estrutural (tanto sintática como discursiva), devido aos diversos modos existentes de se expressar uma mesma mensagem.

[1] Nesta dissertação, os textos-exemplo foram extraídos do corpus TeMário (Pardo e Rino, 2003), disponível para download em: http://www.linguateca.pt/Repositorio/TeMario

Os seguintes passos ilustram a tarefa dos sumarizadores humanos na produção dos sumários: interpretação do texto-fonte, envolvendo leitura e entendimento do texto; identificação das informações mais relevantes; condensação do conteúdo e reescrita do texto, envolvendo nova estruturação e nova expressão lingüística. Esses passos remetem às fases de análise, transformação e síntese que compõem a arquitetura de um sistema de SA profundo proposta por Sparck Jones (1993a), como será apresentada na próxima seção.

2.3 Arquitetura de um Sistema de Sumarização Automática Baseado em Conhecimento Profundo

Do ponto de vista da abordagem profunda, a SA consiste em simular a tarefa de sumarização humana em um processo de identificação das informações mais relevantes de um texto, de modo a permitir a estruturação do sumário com base nessas informações. Entretanto, os principais problemas dessa abordagem estão relacionados à identificação de tais informações e à forma como elas são sintetizadas. Na tentativa de minimizar esses problemas, Sparck Jones (1993a) propõe um modelo que sugere a simulação do próprio processo humano de sumarização, composto por três etapas básicas: a) interpretação do texto-fonte, resultando em uma representação interna, conceitual, abstraída da forma lingüística original; b) geração da representação interna do sumário a partir da representação interna do texto-fonte; c) realização lingüística da representação interna do sumário, produzindo o sumário propriamente dito. Essas três etapas correspondem, assim, aos processos de *Análise*, *Transformação* e *Síntese* que compõem a arquitetura genérica de um sistema de SA baseado em conhecimento profundo, ilustrada na Figura 3.

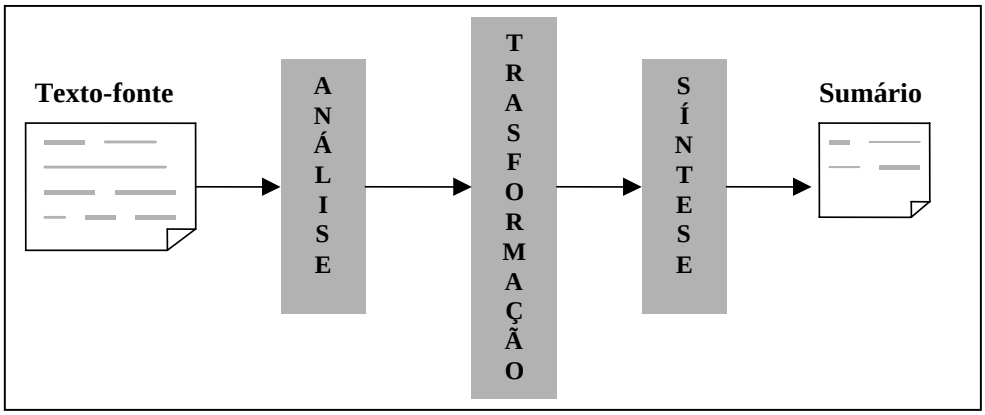

Figura 3: Arquitetura genérica de um sistema de SA

Segundo Sparck Jones, essa arquitetura deve contemplar três tipos de informação: o *lingüístico*, o *informativo* (ou de domínio) e o *comunicativo*, remetendo a questões semânticas e pragmáticas. Assim, é necessário que haja uma linguagem de representação que possibilite o inter-relacionamento entre as unidades discursivas e engenhos de inferência capazes de interpretar o texto-fonte e gerar sua forma condensada correspondente.

Do ponto de vista prático, a análise automática pode fazer uso de um analisador discursivo, para obter a representação conceitual do discurso subjacente ao texto-fonte, através da qual possam ser recuperadas as relações entre os segmentos discursivos, bem como sua relevância para a SA. Assim, a representação conceitual resultante do processo de análise pode servir de base para a transformação em estrutura do sumário. É nesta fase que se concentra este projeto de mestrado, como será mostrado no Capítulo 6. Antes, porém, serão descritas as teorias de estruturação de discurso mais significativas para este trabalho (Capítulos 3 e 4) e também alguns trabalhos de SA que são correlatos a este (Capítulo 5).

3. RHETORICAL STRUCTURE THEORY

A *Rhetorical Structure Theory – RST* (Mann and Thompson, 1987), inicialmente desenvolvida para a análise discursiva de textos e, portanto, para a interpretação, tem sido muito usada na SA.

A RST 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. 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, caso 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. Nesses casos, tem-se uma relação multinuclear, isto é, com mais de um núcleo e nenhum satélite. Assim, as relações RST são divididas em duas classes: hipotáticas e paratáticas (Marcu, 1997a). 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. A Tabela 1[2] apresenta o conjunto de relações retóricas da RST.

[2] As relações adotadas neste trabalho (vide Capítulo 6) são definidas no Apêndice A. As definições daquelas não utilizadas aqui podem ser recuperadas da obra de referência (Mann and Thompson, 1987).

Tabela 1: Conjunto de relações retóricas da RST

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

Considere, por exemplo, as Figura 4 e 5, as quais ilustram relações mononucleares e multinucleares, respectivamente. No texto da Figura 4[3] a *EDU* 1 é o núcleo (N) e a *EDU* 2 é o satélite (S) da relação retórica *PURPOSE*, pois este apresenta uma situação que será realizada pela atividade apresentada no núcleo. Em outras palavras, pode-se dizer que uma das possíveis interpretações dessa estrutura seria "N tem como propósito S", como ilustra o texto correspondente. No texto da Figura 5 a relação retórica *SEQUENCE* indica a seqüência de eventos entre as *EDUs* 1 e 2, sendo que as duas possuem o mesmo grau de importância. Como visto na Tabela 1, o conjunto de relações da RST inclui apenas três relações multinucleares.

[3] A partir daqui, todos os textos-exemplo apresentam as *EDUs* numeradas, para referência.

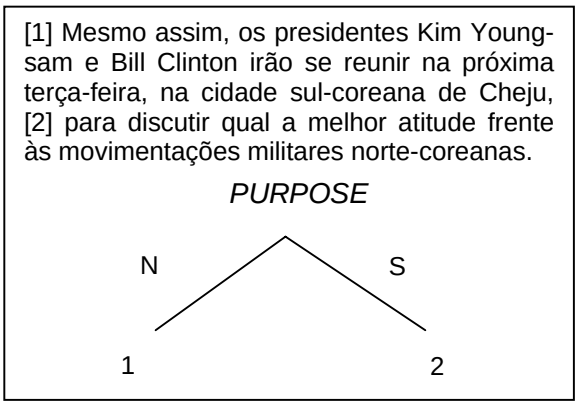

Figura 4: Exemplo de relação mononuclear

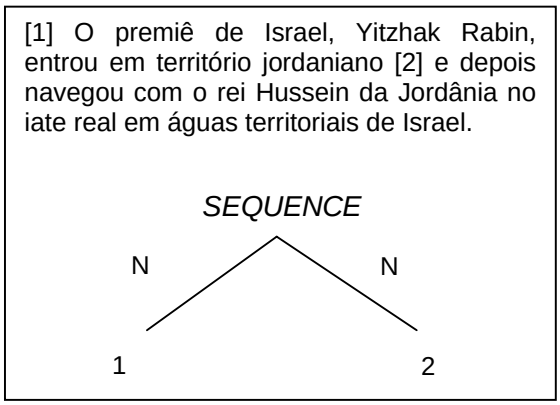

Figura 5: Exemplo de relação multinuclear

A construção de uma estrutura RST de um texto é composicional: relacionam-se duas ou mais unidades elementares como, por exemplo, orações simples, formando-se pequenas subestruturas. Essas subestruturas, por sua vez, podem compor estruturas mais complexas, formadas por segmentos textuais mais elaborados (sentenças, parágrafos, por exemplo). Assim, uma 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 4 e 5.

Ao desenvolver um analisador retórico automático para o inglês, Marcu (1997a) introduziu novas relações retóricas na RST. Dentre elas, foram introduzidas relações para tratar os segmentos encaixados, por exemplo, aqueles introduzidos por orações subordinadas relativas. Essas relações são indicadas por "-e" no final de seu nome (de *embedded*, no inglês).

Como exemplo, considere o texto-exemplo da Figura 6, na qual a *EDU* 2 é o segmento encaixado, aqui relacionado à *EDU* 1 pela relação *ELABORATION-e*. Em geral, as *EDUs* encaixadas seccionam a *EDU* principal em duas partes não contíguas, na superfície textual. Para representar esses casos, Marcu propõe a relação SAME-UNIT, que simplesmente indica essa não contigüidade. É válido dizer que, embora a estruturação retórica seja um fenômeno profundo, a proposta de segmentação de *EDUs* baseia-se na superfície textual, para fins computacionais.

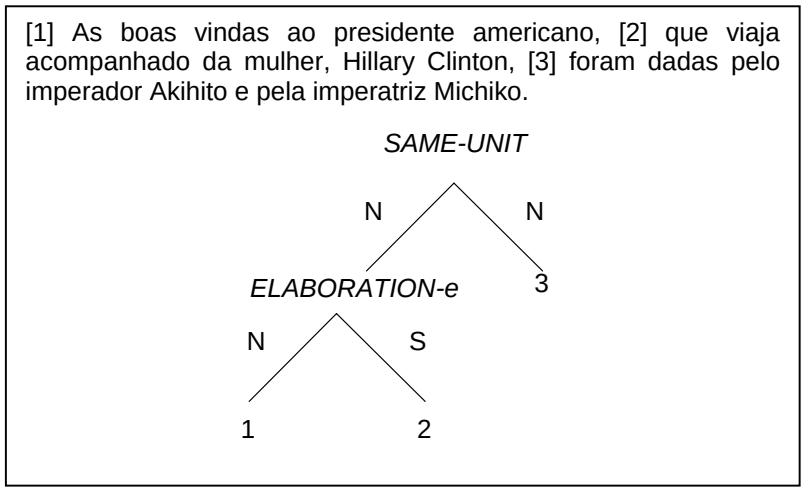

Figura 6: Estrutura retórica com segmento encaixado

A Tabela 3 apresenta o conjunto de relações retóricas introduzidas por Marcu (onde (-e) indica as relações que também podem ser encaixadas). Algumas delas podem ser tanto mononuclear como multinuclear, conforme mostrado na tabela. Várias dessas relações também são consideradas neste trabalho (vide Capítulo 6) e suas definições podem ser recuperadas no Apêndice A[4].

[4] Para as definições do conjunto completo de relações vide Carlson and Marcu (2001).

Tabela 2: Conjunto de relações retóricas introduzidas por Marcu

Relação Retórica	**Mononuclear**	**Multinuclear**
ANALOGY (-e)	X	X
ATRIBUTTION (-e)	X	
CAUSE (-e)	X	
CAUSE-RESULT		X
COMPARISON (-e)	X	X
COMMENT (-e)	X	
COMMENT-TOPIC		X
CONCLUSION (-e)	X	X
CONSEQUENCE (-e)	X	X
CONTINGENCY (-e)	X	
DEFINITION (-e)	X	
ELABORATION-ADDITIONAL (-e)	X	
ELABORATION-PART WHOLE (-e)	X	
ELABORATION-PROCESS-STEP (-e)	X	
ELABORATION-OBJECT-ATTRIBUTE (-e)	X	
ELABORATION-GENERAL-SPECIFIC (-e)	X	
EXAMPLE (-e)	X	
EXPLANATION-ARGUMENTATIVE (-e)	X	
HIPOTETICAL (-e)	X	
INTERPRETATION (-e)	X	X
INVERTED-SEQUENCE		X
LIST		X
MANNER (-e)	X	
PREFERENCE (-e)	X	
PROBLEM-SOLUTION (-e)	X	X
QUESTION-ANSWER (-e)	X	X
REASON (-e)	X	X
RESULT (-e)	X	
RHETORICAL-QUESTION (-e)	X	
SAME-UNIT		X
STATEMENT-RESPONSE (-e)	X	X
TEMPORAL-BEFORE (-e)	X	
TEMPORAL-SAME-TIME (-e)	X	X
TEMPORAL-AFTER (-e)	X	
TEXTUAL-ORGANIZATION		X
TOPIC-COMMENT		X
TOPIC-DRIFT	X	X
TOPIC-SHIFT	X	X

A Figura 8 mostra um exemplo de análise RST mais complexa que os exemplos anteriores para o texto-exemplo 1 (Figura 1), reproduzido na Figura 7 por conveniência. Vale notar que, com exceção das relações *ELABORATION* e *CIRCUMSTANCE*, todas as relações usadas nesse exemplo foram propostas por Marcu.

[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 [5] que vem sendo implantada na empresa desde outubro do ano passado, [6] quando a Pirata se iniciou no Programa Sebrae de Qualidade Total.
[7] Dona de 65% do mercado mineiro de temperos, condimentos e molhos, a Pirata reúne atualmente 220 funcionários. [8] A coordenadora do programa de qualidade na empresa, Márcia Cristina de Oliveira Neto, disse que [9] ainda não é possível dimensionar os ganhos financeiros que "certamente" a empresa terá, em conseqüência da melhoria da qualidade de seus produtos e serviços. [10] Por enquanto, os benefícios mais visíveis, segundo ela, são a organização e a limpeza da fábrica. [11] "Também a relação entre as pessoas tem melhorado bastante. As informações estão mais claras e os funcionários e clientes, mais satisfeitos".

Figura 7: Texto-exemplo 1

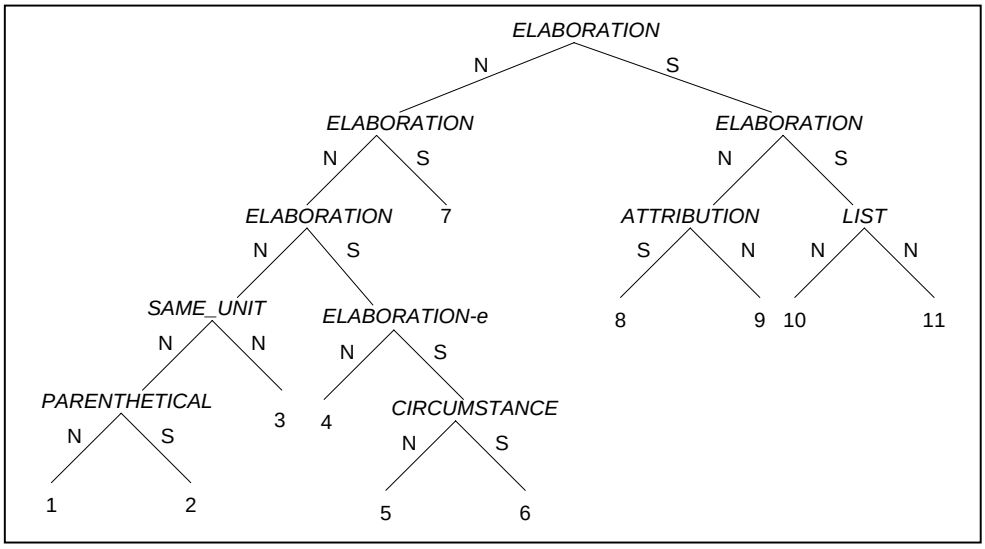

Figura 8: Estrutura RST do texto-exemplo 1

4. VEINS THEORY

A *Veins Theory – VT,* ou Teoria das Veias, (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 (e, portanto, que uma mensagem possa ser recuperada dele) a partir da garantia da coerência local. Como já foi mencionado, suas regras propõem um relacionamento entre a estrutura RST de um texto e suas cadeias de co-referências (CCRs), com base na noção de nuclearidade e na delimitação do domínio de acessibilidade referencial para cada unidade do discurso. A veia de uma unidade é definida como um conjunto de unidades do discurso que podem conter o antecedente de uma anáfora.

Segundo a *Veins Theory*, as veias definidas sobre uma árvore RST são subseqüências da seqüência de unidades elementares que compõem o discurso e são determinadas de acordo com o seguinte algoritmo:

Para todo n $\in$ *ARST*
Se n é um nó folha
então head de n é igual a n
Senão
head de n é igual à concatenação das 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 a sua 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.

Ao aplicar esse algoritmo para o cômputo das veias da árvore retórica da Figura 8 (Capítulo 3), por exemplo, obtêm-se as *heads* (*h*) e veias (*v*) apresentadas em itálico na Figura 9.

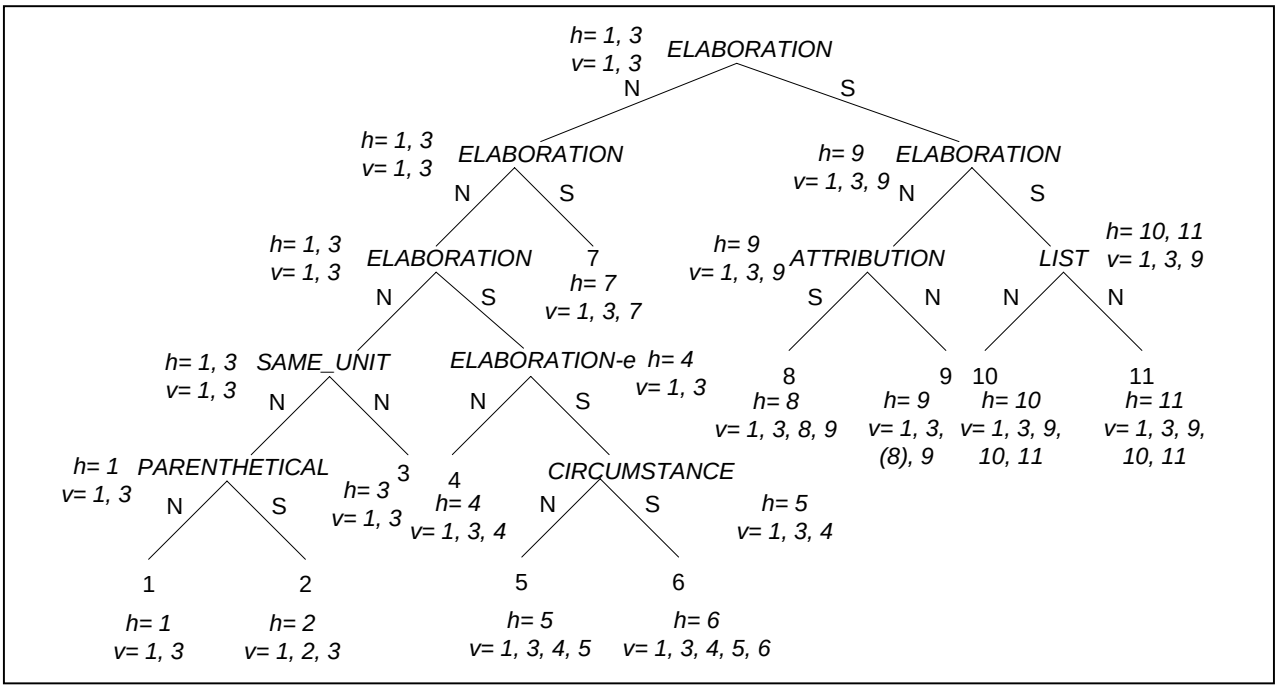

Figura 9: Cômputo das veias para a árvore RST do texto-exemplo1[5]

Para melhor ilustrar o domínio de acessibilidade referencial definido pelo algoritmo das veias, considere a *EDU* [4], reproduzida a seguir, com o termo anafórico ilustrado em negrito.

> [4] Os ganhos são atribuídos pela diretoria d**a fábrica** à nova filosofia.

De acordo com a Figura 9, sua veia é composta pelas *EDUs* [1] e [3]. Assim, o antecedente da anáfora "*a fábrica"* pode estar presente em qualquer uma das *EDUs* [1] ou [3]. Neste exemplo, seu antecedente encontra-se em [1] como se pode constatar no segmento de texto completo, reproduzido 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.
>
> [4] Os ganhos são atribuídos pela diretoria d**a fábrica** à nova filosofia.

[5] Vide Figura 7.

Um sumarizador automático, por exemplo, poderia focar somente na *EDU* [4] para construir um sumário de uma única sentença. No entanto, ao focar somente nessa *EDU*, a coerência do sumário poderia ser perdida, se a anáfora "*a fábrica"* fosse introduzida sem o seu antecedente. Por essa razão, a *EDU* [1] deveria ser incluída e, assim, também a *EDU* [3], já que [1] e [3] constituem a mesma unidade (indicada pela relação *SAME-UNIT* na Figura 9). As heurísticas propostas neste trabalho buscam prevenir justamente esse problema: o de se incluir uma *EDU* que contém uma anáfora em um sumário sem que se inclua a *EDU* que contém o seu antecedente, como será apresentado no Capítulo 6.

5. TRABALHOS CORRELATOS

Neste capítulo são descritos alguns trabalhos de SA que, de alguma forma, se relacionam com o modelo proposto nesta dissertação. Particularmente, o modelo de Marcu (1997a), adotado pelo RHeSumaRST (vide Capítulo 6), também é apresentado.

Dois tipos de trabalhos são destacados aqui: a) aqueles que utilizam a RST para a SA e b) aqueles que contemplam à *Veins Theory* na SA.

5.1 O uso da RST na Sumarização Automática

No contexto da SA, dois fatores tornam a RST interessante: a) a nuclearidade pode corresponder à relevância, isto é, uma unidade nuclear pode fornecer informação mais relevante do que o seu satélite e b) a escolha das unidades discursivas para compor os 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. Nesse caso, o problema da SA consiste em determinar quais satélites, de uma estrutura RST de um texto a ser sumarizado, podem ser retirados sem perda de coerência. As propostas apresentadas nesta seção se baseiam na arquitetura genérica proposta por Sparck Jones (1993a), descrita no Capítulo 2.

5.1.1 Proposta de Marcu.

Marcu (1997a, 1997b, 1999, 2000) propõe um sistema que se baseia na estrutura retórica de textos escritos em inglês para calcular a saliência das unidades elementares do discurso (*EDUs*), para a produção dos sumários.

A cada nó interno da árvore RST de um texto-fonte o sistema atribui um conjunto promocional (*promotion set*) formado pelas unidades salientes ou promocionais daquele nó. Essas unidades salientes são consideradas as mais importantes, no segmento de discurso correspondente, e são determinadas 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 dadas pela união das unidades mais salientes dos filhos nucleares imediatos do referido nó;

Aplicando-se repetidamente o conceito de saliência a cada nó da árvore RST, pode-se obter uma ordem de importância de todas as *EDUs*. A premissa dessa abordagem é que as *EDUs* que estão no conjunto promocional mais próximo da raiz da árvore são mais importantes do que aquelas que se encontram em níveis mais profundos na árvore. Assim, o cômputo da saliência das *EDUs* se baseia tanto na nuclearidade quanto em sua profundidade na estrutura RST e as *EDUs* que estão em conjuntos promocionais mais próximos da raiz têm um *score* (peso) maior do que aquelas que estão em conjuntos promocionais em níveis mais profundos.

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

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

onde:

prom(D) é o conjunto promocional de um nó em D

paren(D) é o conjunto de unidades pais de um nó em D

C(D) é o conjunto de subárvores filhas de um nó em D

Considere a estrutura retórica do texto-exemplo 1, ilustrada na Figura 8 (Capítulo 3). As *EDUs* mais salientes de um dado segmento são apresentadas em negrito na Figura 10. Por exemplo, no segmento de discurso formado pelas *EDUs* 1, 2 e 3 (relação retórica *SAME-UNIT*) as *EDUs* 1 e 3 são as mais saliente, pois são núcleos da relação. Segundo esse modelo de Marcu, o cômputo da saliência de seus nós leva à *EDU(s)* mais saliente(s) de todo o texto. Neste exemplo, as *EDUs* 1 e 3 são indicadas como mais salientes.

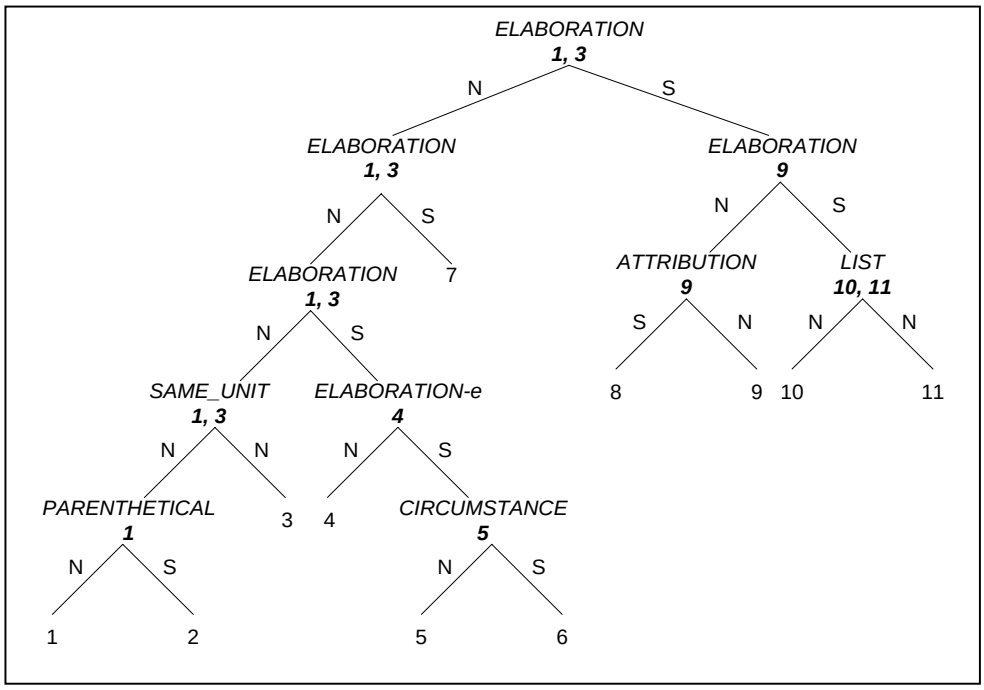

Figura 10: Saliência das EDUs do texto-exemplo 1, segundo a proposta de Marcu

Conforme a Figura 10, a ordem de importância de cada *EDU* é a seguinte: 1, 3> 9 > 7, 10, 11> 4, 8> 5> 2, 6 (x > y indica que x é mais importante do que y; cadeias de *EDUs* separadas por vírgula indicam que elas têm o mesmo grau de importância). Segundo esse modelo, qualquer discurso terá a ordem de importância de suas *EDUs* definida por uma relação do tipo: $C_1 > C_2 > ... > C_n$, para qualquer cadeia de *EDUs* C_i, sendo esta uma única *EDU* ou um conjunto de *EDUs* separadas por vírgula (caso em que elas são igualmente importantes). Desse modo, considerando-se $|C_i|$ como o tamanho de uma cadeia qualquer, dada pelo número de *EDUs* que a compõe, $\sum_{i=1}^{n} |C_i|$ resultará no número de *EDUs* do discurso todo.

Uma vez obtida a ordem de importância das *EDUs*, o sistema pode gerar sumários de tamanhos variados obedecendo tal ordem, de acordo com a taxa de compressão desejada pelo usuário. Por exemplo, se o usuário desejar um sumário bem curto, o sistema pode gerar um sumário contendo apenas as *EDUs* 1 e 3, classificadas como as mais importantes. Caso o usuário queira um sumário um pouco mais longo, o sistema pode incluir também as unidades 9, 7, 10 e 11 e assim por diante. Essa classificação de saliência também é adotada

pelo RHeSumaRST, a fim de obter uma ordem de importância das *EDUs* para a aplicação das heurísticas de poda, como descrito no Capítulo 6.

Para avaliar seu sistema, Marcu (1997b) conduziu um experimento utilizando cinco artigos da revista "*Scientific American*", considerados por ele bem escritos, com tamanhos variando de 161 a 725 palavras. Os textos foram dados a 13 humanos, que atribuíram um *score* variando de 0 a 2 para cada *EDU* de um texto, segundo sua relevância para um potencial sumário. Os textos foram, ainda, analisados por dois lingüistas especialistas em RST, que construíram suas respectivas árvores retóricas. Os mesmos textos foram dados ao sistema que também construiu suas árvores RST e computou a saliência de cada *EDU*.

A distribuição da saliência para cada texto-fonte, gerada automaticamente, foi, então, comparada com a distribuição de referência atribuída pelos humanos. Para a sumarização, foi fixada uma taxa de compressão e foram gerados os sumários a partir das estruturas RST e da distribuição de saliência, tanto para o corpus de referência quanto para o corpus de estruturas RST geradas automaticamente. Enfim, cada par de sumários foi comparado, para se medir o desempenho do sistema que apresentou cobertura de 53% e precisão de 50%.

Adicionalmente, o sistema Microsoft Summarizer do Microsoft Office97 foi usado como *baseline* para comparação com os resultados obtidos pelos analistas. Ao selecionar a mesma porcentagem de unidades consideradas mais importantes pelos analistas, o Microsoft Summarizer obteve cobertura de 28% e precisão de 26%. Assim, Marcu sugere que a estrutura RST de um texto pode ser usada na SA, uma vez que a dicotomia específica da RST entre núcleos e satélites permite determinar as unidades discursivas mais relevantes de um texto.

5.1.2 Proposta de Ono et al.

Apesar de também usar a RST, o modelo de Ono et al. (1994) é muito distinto do modelo de Marcu em relação ao método de condensação da árvore RST para a produção do sumário.

A saliência de cada *EDU* da árvore retórica de um texto é calculada com base na importância relativa das relações retóricas, as quais são classificadas em três tipos:

RightNucleus, LeftNucleus e *BothNucleus*. Nas relações *RightNucleus* o nó direito de um ramo da árvore é considerado mais importante que o nó esquerdo, enquanto que, nas relações *LeftNucleus,* o nó esquerdo é considerado o mais importante. Nas relações *BothNucleus* ambos os nós apresentam o mesmo grau de importância. No que se refere à RST, as relações *RightNucleus* e *LeftNucleus* remetem as relações mononucleares e as relações *BothNucleus* remetem as relações multinucleares.

O sistema construído com base nessa classificação atribui um *score* para cada nó de uma relação retórica segundo seu grau de importância. O *score* 0 é atribuído ao nó mais nuclear, ou seja, o núcleo mais à esquerda da árvore RST, que é considerado o nó mais importante. Os *scores* dos demais nós são calculados somando-se 1 a cada satélite que é encontrado em todos os níveis de profundidade da árvore. Os nós mais importantes apresentam os menores *scores*. Depois, o sistema poda recursivamente os nós que apresentam o maior *score*, resultando na estrutura RST do sumário.

Considere a estrutura retórica do texto-exemplo 1 (Figura 8). Segundo o modelo de Ono et al., as relações retóricas *ELABORATION, PARENTHETICAL* e *CIRCUMSTANCE-e* são classificadas como *LeftNucleus,* enquanto que a relação *ATTRIBUTION* classifica-se como *RightNucleus*. Já as relações *SAME-UNIT* e *LIST* são classificadas como *BothNucleus.* Os scores obtidos para cada *EDU* dessa estrutura RST, aplicando-se esse modelo, são apresentados em itálico na Figura 11. As curvas pontilhadas representam a fronteira entre os diferentes *scores* atribuídos aos nós. Assim, os nós 1 e 3 têm *score* 0, os nós 2, 4, 7 e 9 têm *score* 1, os nós 5, 8, 10 e 11 têm *score* 2 e o nó 6 tem *score* 3. Portanto, tem-se a seguinte ordem de importância das *EDUs*: 1, 3 > 2, 4, 7, 9 > 5, 8, 10, 11 > 6.

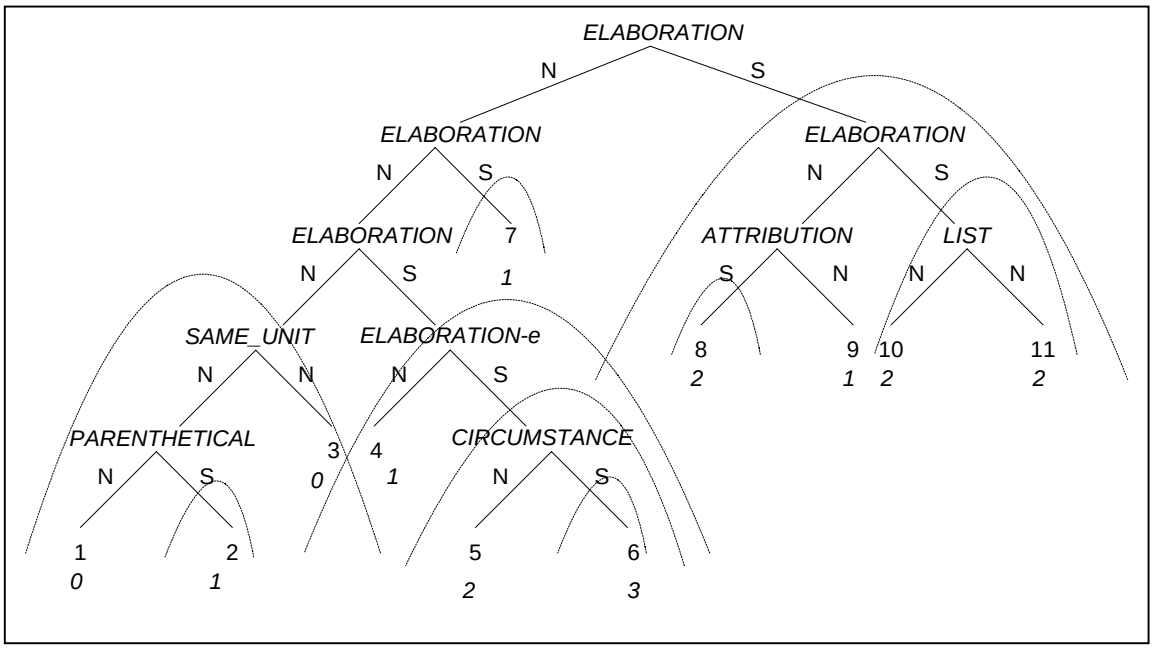

Figura 11: Saliência das EDUs do texto-exemplo 1, segundo a proposta de Ono et al

Assim como no modelo de Marcu, essa ordem de importância permite a construção de sumários com vários graus de granularidade. Um sumário mínimo, por exemplo, deve conter as *EDUs* 1 e 3.

O sistema proposto foi avaliado com base em um corpus de 72 textos, sendo 30 textos de artigos editoriais e 42 artigos técnicos. Os textos foram dados a três humanos, que selecionaram as sentenças-chave de cada um e indicaram as mais importantes dentre elas. Essas sentenças foram usadas como dado de referência para o cálculo de cobertura do sistema, isto é, a razão entre as sentenças-chave indicadas automaticamente e as sentenças-chave indicadas pelos humanos. A Tabela 3 apresenta os resultados obtidos, considerando-se todas as sentenças-chave e somente as sentenças-chave mais importantes.

Tabela 3: Cobertura dos sumários

Textos	**Quantidade**	**Cobertura**	
		Sentenças-chave	**Sentenças-chave mais importantes**
Artigos editoriais	30	41%	60%
Artigos técnicos	42	51%	74%

Ono et al. concluem que a estrutura RST de um texto fornece uma ordem natural de importância entre suas unidades discursivas e, portanto, pode ser usada de maneira eficiente para determinar as informações mais salientes para compor os sumários. Além disso, os autores ainda apontam alguns méritos da RST para a SA. Primeiro, ao contrário dos sumários gerados por sistemas convencionais baseados na freqüência de palavras-chave, os sumários gerados com base na RST são consistentes com o texto-fonte, pois as relações entre as *EDUs* dos sumários refletem as relações do texto-fonte. Segundo, uma vez obtida a estrutura retórica de um texto, sumários de vários tamanhos podem ser gerados. Terceiro, os sistemas baseados na RST podem ser usados para textos de qualquer domínio.

5.1.3 Proposta de O'Donnel

A proposta de O'Donnel (1997) é ainda distinta das duas propostas anteriores em relação ao modo como se calcula a relevância das unidades discursivas para compor o sumário.

Inicialmente, o usuário do sistema (especialista em RST) deve classificar as relações retóricas pelo seu grau de importância, atribuindo-lhes *scores* que variam de 0 a 1.0. Por exemplo, para a relação *ELABORATION* pode ser dado um *score* de 0.40 (baixa relevância), enquanto que para a relação *PURPOSE* pode ser dado um *score* de 0.70, por apresentar informação mais relevante para a satisfação do objetivo do texto do que a relação *ELABORATION*. Esse sistema classificatório pode variar, dependendo do tipo de texto, do domínio ou das intenções comunicativas consideradas.

Com base nos *scores* definidos pelo especialista, o sistema atribui a cada nó da árvore um grau de importância que varia entre 0 e 1.0. O nó mais nuclear, ou seja, o núcleo mais a esquerda da árvore RST, recebe automaticamente o *score* máximo (1.0). O grau de importância dos demais nós é calculado com base no valor do nó mais nuclear mais próximo ao nó em foco, multiplicado pelo *score* da relação retórica que domina o nó mais nuclear.

A Figura 12 mostra a estrutura retórica do texto-exemplo 1 (Figura 8) e os *scores* atribuídos a cada nó (em itálico). *Scores* fictícios foram associados às relações retóricas (*ATTRIBUTION* e *PARENTHETICAL* = 0.20, *ELABORATION, ELABORATION-e* e

SAME-UNIT = 0.40, *LIST* = 0.50 e *CIRCUMSTANCE* = 0.70), resultando na seguinte ordem de importância dos nós: 1, 3 > 4, 6, 8 > 5 > 2 > 9, 10 > 7.

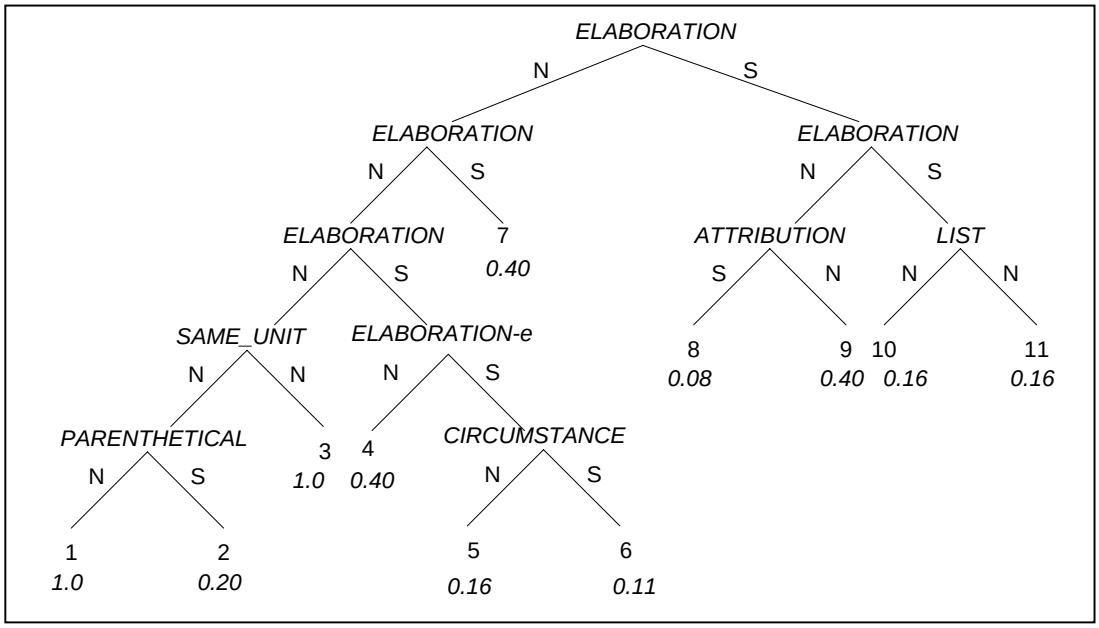

Figura 12: Saliência das *EDUs* do texto-exemplo 1, segundo a proposta de O'Donnel

Analisando-se passo a passo o processo de atribuição de *scores* em cada *EDU* tem-se:

1. Os nós mais nucleares (nó 1 e 3) recebem o valor máximo 1.0;
2. O valor do nó 2 é igual ao valor dos nós 1 e 3, que são os núcleos mais nucleares mais próximos do nó 2, multiplicado pelo *score* da relação *Parenthetical*, ou seja, 1.0 x 0.20 = 0.20;
3. O valor do nó 4 é igual ao valor dos nós 1 e 3, que são os núcleos mais nucleares mais próximos do nó 4, multiplicado pelo *score* da relação *Elaboration*, ou seja, 1.0 x 0.40 = 0.40;
4. O valor do nó 5 é igual ao valor do nó 4, que é o nó mais nuclear mais próximo do nó 5, multiplicado pelo *score* da relação *Elaboration-e*, ou seja, 0.40 x 0.40 = 0.16;
5. O valor do nó 6 é igual ao valor do nó 5, que é o núcleo mais nuclear mais próximo do nó 6, multiplicado pelo *score* da relação *Circumstance*, ou seja, 0.16 x 0.70 = 0.11;

6. O valor do nó 7 é igual ao valor dos nós 1 e 3, que são os núcleos mais nucleares mais próximos do nó 7, multiplicado pelo *score* da relação *Elaboration*, ou seja, 1.0 x 0.40 = 0.40;
7. O valor do nó 9 é igual ao valor dos nós 1 e 3, que são os núcleos mais nucleares mais próximos do nó 9, multiplicado pelo *score* da relação *Elaboration*, ou seja, 1.0 x 0.40 = 0.40;
8. O valor do nó 8 é igual ao valor do nó 9, que é o nó mais nuclear mais próximo do nó 8, multiplicado pelo *score* da relação *Attribution*, ou seja, 0.40 x 0.20 = 0.08;
9. O valor dos nós 10 e 11 é igual ao valor do nó 9, que é o nó mais nuclear mais próximo dos nós 10 e 11, multiplicado pelo *score* da relação *Elaboration*, ou seja, 0.40 x 0.40 = 0.16;

Depois de obter a ordem de importância dos nós da árvore retórica, o sistema poda os nós menos importantes, respeitando a taxa de compressão desejada pelo usuário e a ordem de importância das *EDUs*, a fim de produzir o sumário. Por exemplo, se o usuário deseja um sumário muito comprimido, o sistema pode gerar um sumário contendo somente a(s) *EDU*(s) mais importante, neste caso, as *EDUs* 1 e 3. Caso o usuário deseje um sumário um pouco mais longo, o sistema pode incluir também as *EDUs* 4, 6 e 8, por exemplo, respeitando sempre a ordem de importância das unidades discursivas.

O'Donnel afirma que esta técnica pode produzir resultados ruins em alguns casos, uma vez que a nuclearidade nem sempre reflete a centralidade das informações relevantes, pois o autor do texto pode apresentar informações importantes em lugares do texto retoricamente irrelevantes do ponto de vista da estruturação RST.

5.2 O uso da Veins Theory na Sumarização Automática

Em se tratando de sumarização automática, tem-se conhecimento do uso da *Veins Theory* somente em Cristea et al. (2003; 2005). Nesses trabalhos, a sumarização também é baseada na estrutura do discurso, mas, ao contrário da estrutura RST, as árvores são binárias e não há relações retóricas. As estruturas discursivas são obtidas combinando

restrições impostas com base em marcadores discursivos e em restrições de acessibilidade referencial definidas pela *Veins Theory*. Apesar de não usar a RST, Cristea et al. também consideram o princípio de nuclearidade da RST em seu modelo, para a determinação das veias.

A sumarização de uma estrutura discursiva é realizada com base em um foco específico. Assim, o usuário deve fornecer ao sistema a entidade que o sumário terá como foco como, por exemplo, um objeto ou um personagem do texto. O sistema, então, verifica em qual veia há maior ocorrência da entidade foco e produz o sumário com base nessa veia.

Como ilustração, considere o texto-exemplo 2[6] apresentado na Figura 13 (os marcadores discursivos estão sublinhados) e sua subjacente estrutura discursiva ilustrada na Figura 14 (os núcleos são apresentados em cinza).

> [1] Maria went alone to the market because [2] Simon had to stay at home with the baby. [3] Simon is a good friend of mine and [4] he also helped me in a number of situations. For instance, [5] he was very helpful when [6] I had the problem with the car. [7] I think she has a lot of trust in him to let him alone with the child. [8] You know how Maria is : [9] she is not very hurried to give credit to anybody.

Figura 13: Texto-exemplo 2

[6] Exemplo extraído de Cristea et al. (2003).

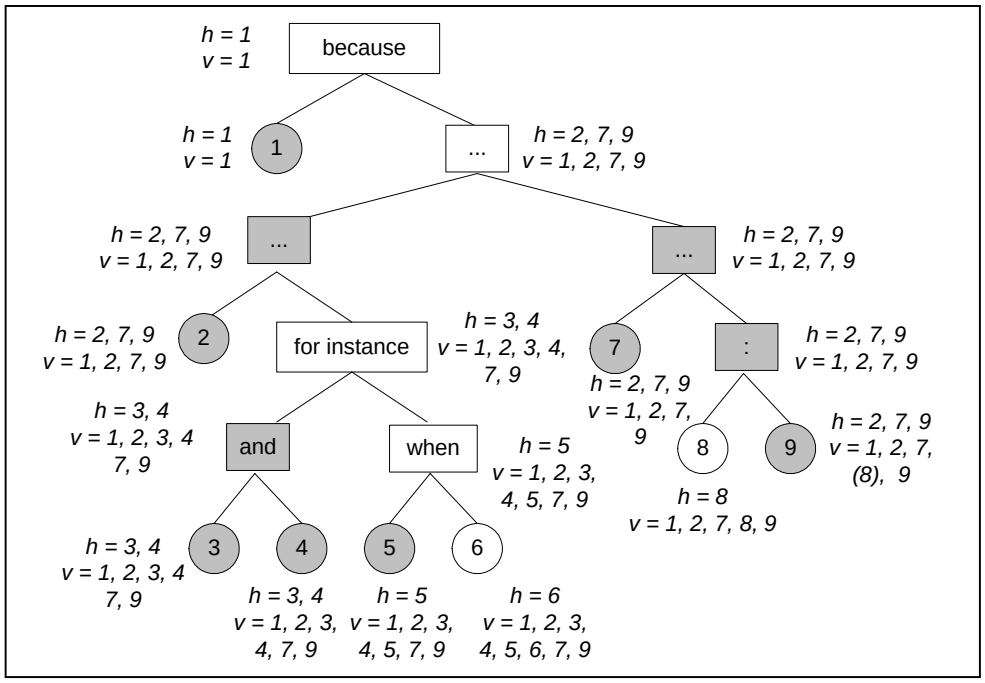

Figura 14: Estrutura discursiva do texto-exemplo 2

Suponha que o usuário deseja um sumário com foco na entidade "Maria". Essa entidade está sendo referenciada nas *EDUs* 1, 7, 8 e 9. De acordo com sua estrutura discursiva (Figura 14), a expressão veia que contém todas essas *EDUs* é a seguinte: 1, 2, 7, 8 e 9 e refere-se à veia das *EDUs* 8 e 9. O sumário produzido com base nessa veia é apresentado na Figura 15.

> [1] Maria went alone to the market because [2] Simon had to stay at home with the baby. [7] I think she has a lot of trust in him to let him alone with the child. [8] You know how Maria is : [9] she is not very hurried to give credit to anybody.

Figura 15: Sumário do texto-exemplo 2 com foco na entidade "Maria"

Com o objetivo de avaliar o método de sumarização baseado na *Veins Theory*, Cristea et al. (2003) utilizaram um texto de uma página extraído do "*The Legends of Mount Olympus*". O texto foi distribuído a 57 estudantes, que indicaram as unidades mais relevantes tendo com foco a entidade "Hefaistos", um personagem do texto. Em seguida, foi construído um sumário composto pelas unidades consideradas mais importantes por

mais da metade dos estudantes. Finalmente, esse sumário foi comparado com o sumário gerado automaticamente, para se medir o desempenho do sistema, que apresentou cobertura de 64.71% e precisão de 73.33%.

5.3 Considerações Sobre as Propostas Apresentadas

As propostas de SA apresentadas neste capítulo referem-se ou à RST ou à *Veins Theory*. As propostas que se baseiam na RST têm em comum a correspondência entre a nuclearidade e a importância de unidades discursivas de um texto-fonte, sugerindo que essa correspondência pode ser explorada para determinar as informações mais relevantes do texto a serem preservadas em seu sumário. Nota-se, nessas propostas, que nenhum tipo de recurso é usado para tratar o relacionamento co-referencial durante a estruturação dos sumários e, portanto, que os sumários produzidos com base nessas abordagens podem ser incoerentes.

Já na proposta que se baseia na *Veins Theory*, o processo de sumarização é baseado em foco específico fornecido pelo usuário e a escolha de informações para compor o sumário é feita com base nas veias. A coerência dos sumários é garantida com base na preservação da veia completa que contém a entidade foco. Um problema observado nessa abordagem é que a preservação da veia completa pode resultar em sumários muito longos.

O sistema RHeSumaRST, proposto neste trabalho, incorpora um modelo cooperativo que adiciona à estruturação RST a proposta de relacionamento co-referencial da *Veins Theory*, visando minimizar os problemas de coerência que podem ser introduzidos pela poda de estruturas RST. O modelo também adiciona a proposta de classificação de saliência de Marcu (1997a), como estratégia preliminar a aplicação das heurísticas de poda. Portanto, o RHeSumaRST ainda tem como técnica central a sumarização automática de estruturas RST, conforme apresentado no próximo capítulo.

6. RHeSumaRST: UM SUMARIZADOR AUTOMÁTICO DE ESTRUTURAS RST

No contexto deste trabalho, a sumarização automática (SA) proposta resume-se à poda de estruturas RST de textos e não de textos escritos em alguma língua natural. A poda de estruturas RST baseia-se em heurísticas que consistem em: a) identificar informações irrelevantes ou menos salientes para exclusão da estrutura RST e b) verificar o relacionamento entre termos anafóricos e seus antecedentes, de modo a garantir a preservação do antecedente de um termo anafórico quando o mesmo for incluso na estrutura 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.

Na seção a seguir apresenta-se a metodologia adotada para a especificação das heurísticas de poda, bem como o elenco de heurísticas para, então, apresentar o sistema RHeSumaRST e seu processo de sumarização de estruturas RST (seção 6.2).

6.1 Especificação de Heurísticas com Base em Corpus

A especificação das heurísticas de poda de estruturas RST baseou-se na análise de corpus, com dois propósitos principais: a) identificar construções discursivas cujos satélites apresentassem informações supérfluas e b) verificar os contextos co-referenciais que poderiam introduzir quebras de co-referência na SA. Nesta seção, apresentam-se o corpus escolhido para análise, as etapas de preparação do corpus e a análise de corpus, propriamente dita, para a especificação das heurísticas.

6.1.1 Eleição do Corpus

O corpus escolhido para análise é composto de 30 textos do gênero jornalístico (com um total de 16.370 palavras, aproximadamente 1 e ½ página cada texto), os quais foram extraídos do corpus TeMário (Pardo e Rino, 2003). O número limitado de textos

deveu-se à falta, na época, de um analisador retórico automático e de uma ferramenta de anotação automática de cadeias de co-referências (CCRs) para o português. Outros fatores que também levaram a essa limitação foram a complexidade envolvida nos processos de análise retórica e anotação de CCRs e o tempo despendido para a realização manual desses processos.

A escolha por textos do gênero jornalístico deveu-se a sua linguagem bastante abrangente e de fácil compreensão, facilitando, assim, a análise de discurso. Outra razão para a escolha desse corpus é o fato de que ele é associado também a uma coleção de sumários redigidos por um especialista humano, os quais poderiam ser usados para comparação com os sumários produzidos pelo RHeSumaRST.

6.1.2 Preparação do Corpus

A preparação do corpus para análise consistiu em duas etapas distintas: a) construção das estruturas RST e b) anotação das cadeias de co-referências de cada texto do corpus. As subseções 6.1.2.1 e 6.1.2.2 descrevem cada uma dessas etapas, respectivamente.

6.1.2.1 Análise Retórica do Corpus

Nesta etapa, cada texto do corpus foi anotado retoricamente com o auxílio da ferramenta de estruturação retórica *RST Annotation Tool*[7]. Essa ferramenta fornece apenas um suporte gráfico para a construção e manipulação de árvores retóricas de textos, sendo, portanto, necessário o conhecimento prévio do analista sobre a RST e técnicas de análise de discurso. A análise também contou com o auxílio da ferramenta RhetDB[8], para acesso e manipulação de uma base de dados que contém informações sobre a análise discursiva. A RhetDB incorpora, assim, as estruturas RST construídas com o auxílio da *RST Annotation Tool*, para que o analista de discurso possa armazenar todo o conhecimento relacionado à sua análise.

Para essa anotação, foram utilizadas algumas relações retóricas da RST (Mann & Thompson, 1987) e também algumas relações propostas por Carlson and Marcu (2001).

[7] Disponível em: http://www.isi.edu/~marcu/discourse/AnnotationSoftware.html (último acesso: junho/2005).
[8] Desenvolvida por Thiago A. S. Pardo no contexto de seu projeto de doutorado (Pardo, 2005).

Essas relações são apresentadas na subseção 6.1.2.1.3. Antes, porém, apresentam-se a forma como os textos foram segmentados para análise e a estratégia de análise adotada.

6.1.2.1.1 Segmentação Textual

Para a análise retórica, um texto pode ser segmentado com diversas granularidades, por exemplo, em parágrafos, sentenças, orações, dentre outros. Para segmentar os textos do corpus adotaram-se orações como *EDUs*, seguindo basicamente a proposta de (Carlson and Marcu, 2001). Em alguns casos ocorreram exceções devido à natureza de algumas relações retóricas envolvidas e que nem sempre relacionam orações como é o caso, por exemplo, das relações temporais (*TEMPORAL-AFTER* e *TEMPORAL-SAME-TIME*) e da relação *PARENTHETICAL*. Embora as regras de segmentação de discurso dessa proposta sejam fortemente dependentes da sintaxe, elas se têm mostrado consistentes, havendo sido aplicadas a conjuntos expressivos de textos (tanto em inglês quanto em português) de forma coerente e não ambígua. Marcadores sintáticos e discursivos foram usados para determinar as *EDUs*. Alguns exemplos dessas regras são dados a seguir e foram extraídos de Carlson and Marcu (2001), páginas 26-41[9]:

- Orações principais são consideradas *EDUs*;
- Orações sinalizadas por marcadores discursivos fortes como, por exemplo, *Porque, Apesar de, Conforme, Segundo, Em conseqüência de,* entre outros, são consideradas *EDUs*;
- Orações subordinadas introduzidas por marcadores discursivos são consideradas *EDUs*;
- Orações complementares não são consideradas *EDUs*, exceto quando introduzirem complemento de um verbo de atribuição. Por exemplo: *[1] A companhia disse que [2] fechará a fábrica.*
- Orações coordenadas são consideradas *EDUs* distintas;
- Orações subordinadas substantivas e objetivas não são consideradas *EDUs*;
- Orações relativas, apositivas e parênteses são consideradas *EDUs* encaixadas.

[9] Tradução nossa.

É válido ressaltar a importância de se considerar *EDUs* encaixadas na análise retórica de um texto. Considerando que um texto não é uma simples seqüência de sentenças desconexas, mas sim uma seqüência coerente de enunciados, ele pode conter muitas *EDUs* encaixadas, as quais são responsáveis por manter sua coerência. Portanto, o fato de não considerar as *EDUs* encaixadas na análise, como propõe a RST, implica uma perda considerável de granularidade na estruturação retórica de qualquer texto (Carlson and Marcu, 2001).

A estratégia de análise retórica usada na análise do corpus considera, assim, as *EDUs* encaixadas. Essa estratégia é descrita a seguir.

6.1.2.1.2 Estratégia de Análise Retórica

Segundo Carlson and Marcu (2001), há várias estratégias de análise retórica. Por exemplo, pode-se fazer uma análise incremental, isto é, relacionar primeiramente duas *EDUs*, resultando em uma subestrutura RST, a qual, por sua vez, será relacionada a outra *EDU*. Sucessivamente, a análise incremental resulta, assim, na agregação, uma a uma, de *EDUs* às sub-estruturas em formação. Pode-se, ainda, montar as subestruturas de cada parágrafo do texto isoladamente e depois integrá-los formando uma única estrutura RST completa do texto.

A estratégia de análise usada para anotar os textos do corpus foi a seguinte: em primeiro lugar relacionou-se retoricamente todas as *EDUs* presentes em uma sentença; depois, relacionou-se todas as sentenças de um parágrafo; por fim, todos os parágrafos do texto foram relacionados, formando uma única árvore de estrutura retórica. A estratégia adotada mostrou-se adequada e consistente para quase toda a análise do corpus. Em alguns casos, devido a diferentes estilos de escrita e problemas nas estruturas dos textos, houve necessidade de análise incremental. Para um exemplo detalhado a passo a passo de análise retórica de um texto vide Seno e Rino (2004).

6.1.2.1.3 Conjunto de Relações Retóricas

Inicialmente, utilizaram-se algumas relações retóricas do conjunto original (Mann and Thompson, 1987). No decorrer da análise, percebeu-se a necessidade de outras relações

não contempladas nesse conjunto, as quais foram extraídas do conjunto definido por Carlson e Marcu. A Tabela 4 mostra o conjunto completo das relações utilizadas. Essas relações são definidas no Apêndice A.

Tabela 4: Conjunto de relações retóricas usado na análise retórica do corpus

Relações Retóricas	**Tipo de Relação**
ATTRIBUTION	Mononuclear
CAUSE	Mononuclear
CAUSE-e	Mononuclear
CIRCUMSTANCE-e	Mononuclear
COMPARISON	Mononuclear
COMPARISON-e	Mononuclear
ELABORATION-e	Mononuclear
EXAMPLE	Mononuclear
EXPLANATION-ARGUMENTATIVE	Mononuclear
EXPLANATION-ARGUMENTATIVE-e	Mononuclear
JUSTIFY-e	Mononuclear
LIST	Multinuclear
MEANS	Mononuclear
MEANS-e	Mononuclear
PARENTHETICAL	Mononuclear
PURPOSE-e	Mononuclear
REASON	Mononuclear
REASON-e	Mononuclear
RESULT	Mononuclear
SAME-UNIT	Multinuclear
SUMMARY	Mononuclear
SUMMARY-e	Mononuclear
TEMPORAL-AFTER	Mononuclear
TEMPORAL-SAME-TIME	Mononuclear
CIRCUMSTANCE	Mononuclear
CONCESSION	Mononuclear
CONDITION	Mononuclear
CONTRAST	Multinuclear
ELABORATION	Mononuclear
EVIDENCE	Mononuclear
INTERPRETATION	Mononuclear
JOINT	Multinuclear
JUSTIFY	Mononuclear
PURPOSE	Mononuclear
SEQUENCE	Multinuclear

A seguir, apresenta-se uma síntese da ocorrência de cada relação no corpus.

6.1.2.1.4 Síntese da Análise do Corpus

A tabela 5 mostra o número de ocorrências e a freqüência de cada relação retórica no corpus. Como se pode notar, algumas relações ocorreram com pouquíssima freqüência. Já a relação *ELABORATION* foi a mais freqüente. Isto talvez se justifique pela natureza do corpus: tratando-se de textos jornalísticos, elaborações sobre um mesmo tópico parecem ser mais freqüentes do que em textos de outro gênero.

Tabela 5: Número de ocorrências e freqüência das relações retóricas

Relações Retóricas	Ocorrências	Freqüência (%)
ELABORATION	413	27.9
LIST	314	21.2
ELABORATION-e	140	9.5
ATTRIBUTION	113	7.6
EVIDENCE	105	7.1
SAME-UNIT	89	6.0
SEQUENCE	74	5.0
REASON	45	3.0
CONTRAST	22	1.5
PURPOSE	17	1.2
JUSTIFY	14	0.9
JOINT	14	0.9
PARENTHETICAL	13	0.9
CONDITION	13	0.9
CAUSE	12	0.8
COMPARISON	9	0.6
CIRCUMSTANCE-e	8	0.5
EXPLANATION-ARGUMENTATIVE	7	0.5
CONCESSION	6	0.4
EXPLANATION-ARGUMENTATIVE-e	6	0.4
PURPOSE-e	6	0.4
REASON-e	6	0.4
EXAMPLE	4	0.3
RESULT	4	0.3
TEMPORAL-AFTER	4	0.3
COMPARISON-e	4	0.3
CIRCUMSTANCE	3	0.2
MEANS-e	3	0.2
INTERPRETATION	2	0.1
MEANS	2	0.1
JUSTIFY-e	2	0.1
TEMPORAL-SAME-TIME	1	0.1
SUMMARY	1	0.1
SUMMARY-e	1	0.1
CAUSE-e	1	0.1
Totais de ocorrências	1478	

6.1.2.2 Anotação das CCRs do Corpus

Nesta etapa, cada texto do corpus foi anotado com as cadeias de co-referências (CCRs), tendo como apoio a ferramenta de anotação de co-referências MMAX (Müller and Strube, 2001)[10]. Essa ferramenta fornece apenas um suporte gráfico para a anotação e manipulação de CCRs em textos, sendo necessário, portanto, o conhecimento prévio do especialista anotador, sobre o fenômeno de co-referenciação em questão.

Para a anotação com a MMAX, cada texto deve seguir uma estrutura padrão estabelecida no formato XML. Nesse formato o texto é representado por palavras (*words*), onde cada palavra possui um identificador (id), como exemplificado na Figura 16.

```
…
    <word id="word_39">A</word>
    <word id="word_40">empresa</word>
    <word id="word_41">Produtos_Pirata_Indústria</word>
    <word id="word_42">e</word>
    <word id="word_43">Comércio_Ltda</word>
    <word id="word_44">.</word>
    <word id="word_45">de</word>
    <word id="word_46">Contagem</word>
...
```

Figura 16: Formato do arquivo de entrada da MMAX

Para obter esse formato, cada texto do corpus foi previamente processado com o parser PALAVRAS (Bick, 2000). Após o processamento sintático, a ferramenta Extractor[11] foi usada para gerar o arquivo de palavras. Essa etapa de pré-processamento do corpus foi realizada por Coelho (2004), que também utilizou esse corpus no âmbito de seu projeto de iniciação científica.

Posteriormente, cada texto (isto é, cada arquivo de palavras) foi anotado. O resultado do processo de anotação de um texto na MMAX é um arquivo em XML, conforme o apresentado na Figura 17. Cada anáfora anotada é representada por um elemento <markable>, cujo atributo *span* indica as palavras que a formam, o atributo *pointer* indica o identificador do seu antecedente e o atributo *form* indica qual o seu tipo (por exemplo, se é um pronome, uma descrição definida, etc.).

[10] Disponível em: http://www.eml.org/english/Research/NLP/Downloads (último acesso: junho/2005).
[11] Disponível em: http://abc.di.uevora.pt/visl/main (último acesso: junho/2005).

```
…
    <markable id="markable_64" span="word_471..word_472"
        pointer="markable_1"  form="defNP" />
    <markable id="markable_63" span="word_468..word_472"
        pointer="markable_90" form="defNP" />
    <markable id="markable_62" span="word_457..word_458"
        pointer="markable_53" form="defNP" />
…
```

Figura 17: Formato do arquivo de saída da MMAX

Somente as CCRs do tipo descrições definidas foram consideradas nessa anotação, isto é, aquelas formadas por um sintagma nominal iniciado por um artigo definido (por exemplo, *o escritor Monteiro Lobato, o aeroporto de Cumbica, a cidade,* etc.). Um estudo realizado por Coelho (2004) mostra que 30% dos sumários automáticos gerados pelo sistema GistSumm (Pardo et al., 2003) apresentam quebras de co-referências introduzidas por descrições definidas. Outros trabalhos (por exemplo, Salmon-Alt and Vieira (2002); Vieira et al. (2002)) apontam um alto índice de ocorrência de descrições definidas em corpus de textos jornalísticos. Assim, optou-se por verificar somente a ocorrência desse fenômeno neste trabalho.

A tarefa de anotação das descrições definidas consistiu em identificar as anáforas definidas e relacioná-las a seus respectivos antecedentes. Nenhuma classificação foi atribuída a essa relação, ou seja, se a anáfora é direta, indireta, associativa, etc., uma vez que, para a SA, essa classificação não é relevante. O que importa é se a anáfora pode ou não apresentar problemas de coerência no RHeSumaRST. Várias construções gramaticais foram consideradas nessa anotação, por exemplo, substantivos, sintagmas nominais, entre outros. No entanto, impôs-se como limite máximo construções gramaticais de até uma frase. Isto significa que os termos anafóricos maiores que uma frase não foram anotados. No total, 896 cadeias de co-referências foram anotadas no corpus (em média, 30 ocorrências em cada texto).

Após a preparação do corpus, realizou-se sua análise, propriamente dita, para a especificação das heurísticas, conforme descrito na seção seguinte.

6.1.3 Análise de Corpus

A análise do corpus visou à especificação de heurísticas de poda de estruturas RST de textos-fonte, considerando as restrições fundamentais de exclusão de informações irrelevantes e de preservação da coerência. Assim, as heurísticas de sumarização de estruturas RST devem contemplar esses dois aspectos de naturezas distintas: o primeiro, relacionado à informatividade do sumário; o segundo, à sua coerência. Esses dois aspectos são descritos a seguir.

6.1.3.1 Análise com Foco na Informatividade

Primeiramente, compararam-se as estruturas RST de cada texto do corpus com seus correspondentes sumários manuais (SMs), construídos pelo profissional humano, como mencionado na seção 6.1.1. 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 (vide Mani (2001)).

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 6. A quarta coluna (Freqüência) indica a representatividade do satélite da relação em foco, com base nos SMs.

Tabela 6: Representatividade dos satélites preservados nos SMs

Relação Retórica	**Ocorrência no Corpus**	**Satélites Preservados no SMs**	**Freqüência (%)**
EXPLANATION 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. Portanto, todas essas relações são consideradas nas heurísticas de poda, como apresentado na seção 6.1.2.

Além das relações incluídas na Tabela 6, relações multinucleares também ocorrem no corpus. Entretanto, para a sumarização, elas não são significativas, pois se algum de seus núcleos for incluído no sumário, todos os outros também serão, por apresentarem igual significância. Por essa razão, não há heurísticas para essas relações.

Devido ao tamanho limitado do corpus, buscaram-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. Desse modo, os resultados apresentados na literatura confirmam os resultados da análise.

Como este trabalho também é de natureza discursiva (vide Capítulo 4), mais especificamente, visando evitar a quebra das cadeias de co-referências (CCRs), tarefas analíticas adicionais foram necessárias, como mostra a seção a seguir.

6.1.3.2 Análise com Foco na Coerência

Visando, agora, buscar subsídios para que as heurísticas de poda levassem a sumários coerentes, 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 (Capítulo 4).

Dessa forma, delimitaram-se as veias para cada uma das 30 estruturas RST dos textos do corpus. Após a delimitação das veias, analisou-se, para cada CCR (somente as descrições definidas) 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. Isso indica que 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 (Seno e Rino, 2005a), descrito na seção a seguir.

6.1.4 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 interfiram na coerência, isto é, as *EDUs* a serem excluídas serão somente aquelas que estão fora do domínio de acessibilidade referencial de outras *EDUs* candidatas a inclusão. De um modo geral, cada heurística se baseia na verificação da acessibilidade referencial de cada *EDU* candidata ao sumário, para a exclusão de uma *EDU* satélite. Genericamente, as heurísticas são representadas por uma única regra condicional, definida a seguir:

Seja $T=\{t_1...t_N\}$ um conjunto de *EDUs* que compõem uma estrutura RST de um texto-fonte, $S=\{s_1...s_M\}$ um conjunto que contém somente as *EDUs* de T que são candidatas a compor o sumário do texto (para $M<N$) e $V=\{v_1...v_L\}$ um conjunto de *EDUs* que constituem a veia de uma *EDU* s_i pertencente a S, para $1 \leq i \leq M$.

Se t_i (com $1 \leq i \leq N$) é o satélite de uma relação retórica R qualquer e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.

Para ilustração, apresentam-se, a seguir, para cada relação retórica do corpus, suas respectivas heurísticas de poda, juntamente com sua descrição funcional e um exemplo (os

satélites são apresentados em negrito)[12]. Para simplificação, supõe-se, nesses exemplos, que os satélites não pertencem às veias de outras *EDUs*.

H1. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *ATTRIBUTION* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *ATTRIBUTION* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *CAUSE* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *CAUSE* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *CAUSE-e* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *CAUSE-e* se ela não estiver na veia de uma *EDU* candidata a compor o sumário.
Exemplo: Devido a mudanças legais, facilitaram-se as aposentadorias de professores, **o que fez aumentar a despesa do ministério com inativos.**

H4. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *CIRCUMSTANCE* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *CIRCUMSTANCE* se ela não estiver na veia de uma *EDU* candidata a compor 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.**

[12] Todos os exemplos foram extraídos do corpus escolhido.

H5. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *CIRCUMSTANCE-e* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.

Função: Excluir uma *EDU* satélite de uma relação *CIRCUMSTANCE-e* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *COMPARISON* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.

Função: Excluir uma *EDU* satélite de uma relação *COMPARISON s*e ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *COMPARISON-e* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.

Função: Excluir uma *EDU* satélite de uma relação *COMPARISON-e s*e ela não estiver na veia de uma *EDU* candidata a compor o sumário.

Exemplo: Como era de se esperar, a municipalização encontrou e encontra muitas resistências - **comparáveis às que foram encontradas no programa de privatização.**

H8. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *CONCESSION* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.

Função: Excluir uma *EDU* satélite de uma relação *CONCESSION s*e ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *CONDITION* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.

Função: Excluir uma *EDU* satélite de uma relação *CONDITION s*e ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *ELABORATION* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *ELABORATION* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *ELABORATION-e* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *ELABORATION-e* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *EVIDENCE* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *EVIDENCE* se ela não estiver na veia de uma *EDU* candidata a compor o sumário.
Exemplo: Como conseqüência, os funcionários da Pirata se tornaram mais exigentes. **A primeira exigência foi à implantação de algum tipo de lazer na hora do almoço.**

H13. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *EXAMPLE* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *EXAMPLE* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *EXPLANATION-ARGUMENTATIVE* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.

Função: Excluir uma *EDU* satélite de uma relação *EXPLANATION-ARGUMENTATIVE* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *EXPLANATION-ARGUMENTATIVE-e* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *EXPLANATION-ARGUMENTATIVE-e* se ela não estiver na veia de uma *EDU* candidata a compor o sumário.
Exemplo: Em outubro cai um pilar da apartheid - **a lei que dividia locais públicos entre brancos e negros.**

H16. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *INTERPRETATION* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *INTERPRETATION* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *JUSTIFY* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *JUSTIFY* se ela não estiver na veia de uma *EDU* candidata a compor o sumário.
Exemplo: Pode-se evitar tudo isso com a experiência, preferivelmente, na atividade a ser exercida. **Experiência é fundamental para minimizar desacertos.**

H18. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *JUSTIFY-e* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *JUSTIFY-e* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *MEANS* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *MEANS* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *MEANS-e* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *MEANS-e* se ela não estiver na veia de uma *EDU* candidata a compor o sumário.
Exemplo: A idéia do relator da revisão constitucional era tão-somente flexibilizar os monopólios, **que seria realizada através de concessão de serviços.**

H21. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *PARENTHETICAL* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *PARENTHETICAL* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *PURPOSE* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *PURPOSE* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *PURPOSE-e* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.

Função: Excluir uma *EDU* satélite de uma relação *PURPOSE-e s*e ela não estiver na veia de uma *EDU* candidata a compor o sumário.
Exemplo: Depois, parte para São Petersburgo e Moscou, **onde participará da cúpula do G-7 (grupo dos sete países mais ricos).**

H24. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *REASON* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *REASON s*e ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *REASON-e* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *REASON-e s*e ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *RESULT* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *RESULT s*e ela não estiver na veia de uma *EDU* candidata a compor o sumário.
Exemplo: **Por causa do mau tempo** não foi possível atingir as armas.

H27. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *SUMMARY* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *SUMMARY s*e ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *SUMMARY-e* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *SUMMARY-e* se ela não estiver na veia de uma *EDU* candidata a compor 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. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *TEMPORAL-AFTER* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *TEMPORAL-AFTER* se ela não estiver na veia de uma *EDU* candidata a compor o sumário.
Exemplo: Foi à primeira ação da Otan contra sérvios **desde o ataque aéreo às suas posições no encrave de Gorazde, em abril.**

H30. Se $t_i \in T$ (com $1 \leq i \leq N$) é o satélite de uma relação retórica *TEMPORAL-SAME-TIME* e $t_i \notin V$ de uma $s_j \in S$ (com $1 \leq j \leq M$), então exclua t_i.
Função: Excluir uma *EDU* satélite de uma relação *TEMPORAL-SAME-TIME* se ela não estiver na veia de uma *EDU* candidata a compor 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.**

6.2. A Sumarização Automática de Estruturas RST

Esta seção apresenta a arquitetura do RHeSumaRST, construído para a aplicação das heurísticas e o seu principal processo: o mecanismo de poda de estruturas RST.

6.2.1 Arquitetura do RHeSumaRST

A Figura 18 ilustra a arquitetura do RHeSumaRST composta de três módulos de processamento em *pipeline*. No primeiro, a estrutura RST do texto-fonte, obtida com a *RST Annotation Tool* (formato SGML), é anotada aplicando-se o algoritmo de delimitação de

veias de Cristea et al. (1998). No segundo, as *EDUs* da estrutura RST são classificadas com base na função de saliência de Marcu (1997) (vide seção 5.1.2). Finalmente, a estrutura RST com as *EDUs* classificadas é podada com base nas heurísticas de poda, resultando no sumário em língua natural. Embora o processo de realização lingüística seja de suprema importância para um sumarizador profundo, como é o caso do RHeSumaRST, esse processo não foi contemplado neste trabalho. O foco do trabalho atual está somente na estruturação discursiva do sumário segundo as duas naturezas previstas na seção 6.1.1.3. Assim, a produção do sumário em língua natural, a partir de sua estrutura RST, se dá simplesmente pela justaposição das *EDUs* na ordem em que ocorrem na árvore RST.

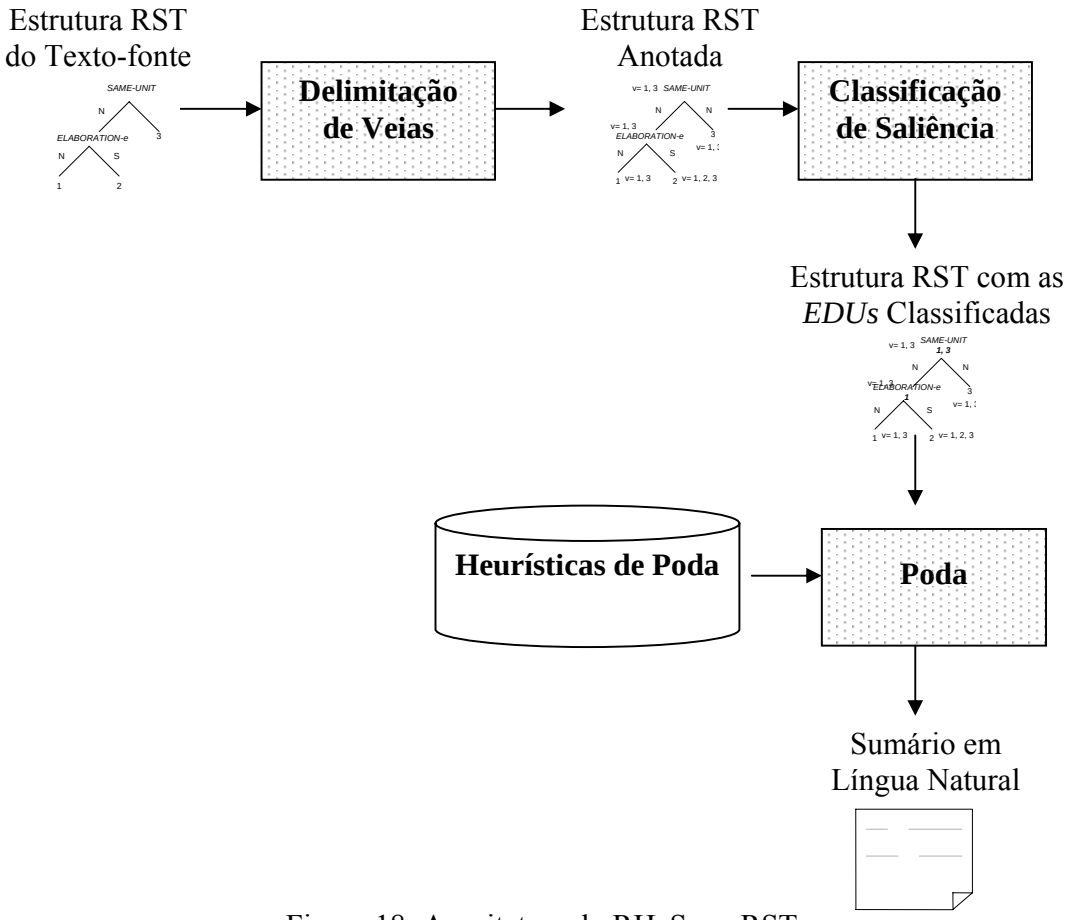

Figura 18: Arquitetura do RHeSumaRST

6.2.2 Processo de Poda do RHeSumaRST

O processo de poda do RHeSumaRST é responsável por podar as informações irrelevantes em uma estrutura RST de um texto para a produção do seu sumário em língua natural. Esse processo se realiza da seguinte maneira: verifica-se para cada *EDU* da estrutura RST, respeitando sua ordem de importância (ou seja, iniciando-se pelas *EDUs* menos importantes), se há alguma heurística que seja aplicável. Caso haja alguma, ela é aplicada e a relação retórica envolvendo-a é excluída, para a reestruturação da árvore RST do sumário. Esse processo se repete até que se atinja a taxa de compressão previamente estabelecida pelo usuário do sistema.

Embora se trate de um sumarizador profundo, 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, só possível calcular o número de unidades elementares (e não de palavras), sendo que em um sumarizador automático completo o tamanho real do sumário final não pode ser delineado pelo módulo de poda, mas somente pelo realizador lingüístico.

A restrição fundamental, aqui, é que a preservação de uma *EDU* qualquer na estrutura do sumário implica a preservação de todas as *EDUs* que compõem a sua veia, mesmo que, em alguns casos, isso provoque um número maior de *EDUs* no sumário final do que o estabelecido pela taxa de compressão.

Finalmente, as *EDUs* da estrutura RST do sumário são justapostas, resultando, assim, no sumário propriamente dito. Para melhor ilustrar esse processo, a subseção a seguir apresenta um exemplo de poda da estrutura RST do texto-exemplo 1 (Figura 7, Capítulo 3).

6.2.2.1 Ilustração do Processo de Poda

Considere a árvore RST do texto-exemplo 1 reproduzida na Figura 19, por conveniência. Após o processo de classificação de *EDUs* com base na função de saliência, obtém-se a seguinte ordem de importância das *EDUs*: 1, 3> 9 > 7, 10, 11> 4, 8> 5> 2, 6 (a(s) *EDU*(s) mais saliente(s) de um dado segmento é apresentada em negrito na figura).

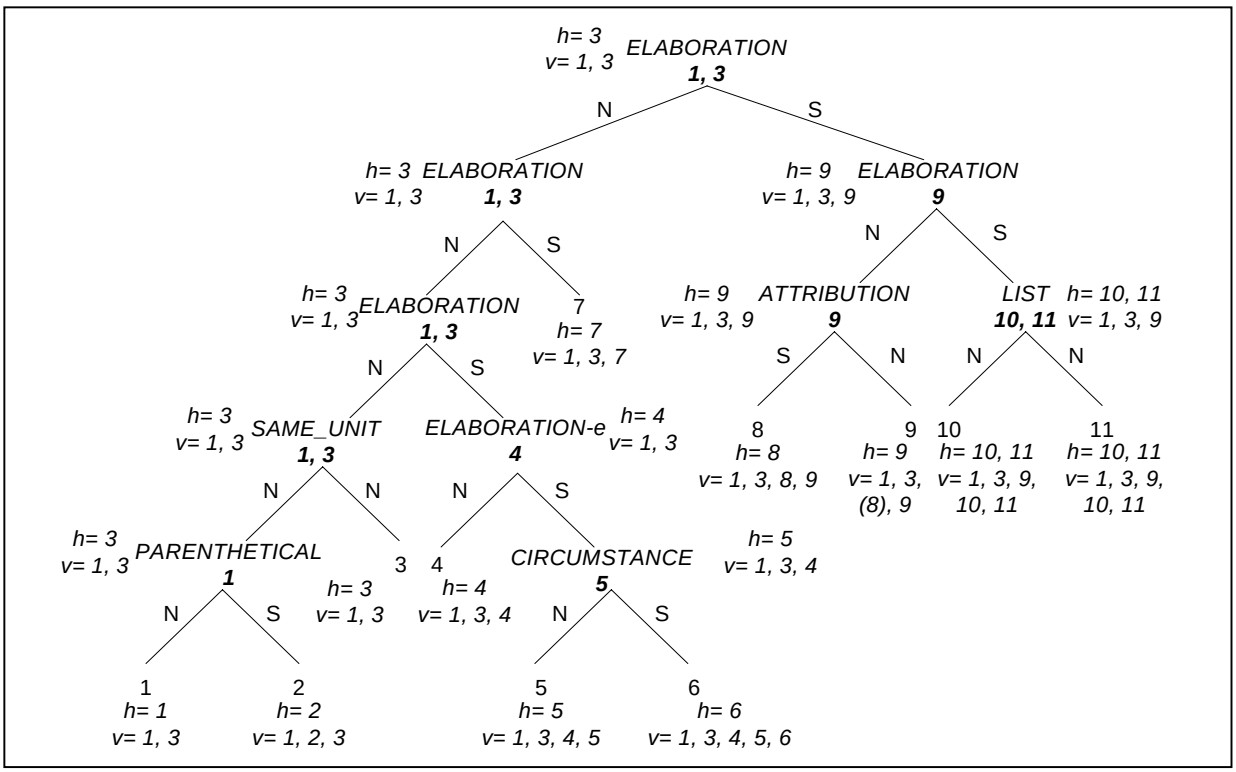

Figura 19: Classificação das EDUs da árvore RST do texto-exemplo 1

Suponha uma taxa de compressão de 70%, ou seja, um sumário composto de 30% das *EDUs* do texto-fonte. Neste exemplo, mais especificamente, composto por 3 *EDUs*. Com base na taxa de compressão e na ordem de importância das *EDUs*, o processo de poda verifica para cada uma delas se há alguma heurística que seja aplicável. É válido lembrar que uma heurística somente se aplica a uma *EDU* se ela não estiver na veia de outras *EDUs* candidatas ao sumário[13].

Conforme a Figura 19, apenas para as *EDUs* 6, 2 e 7 há heurísticas que se aplicam. São elas: H4, H21 e H10, respectivamente, as quais remetem às relações retóricas *CIRCUMSTANCE, PARENTHETICAL* e *ELABORATION*. Após a aplicação dessas heurísticas, essas relações são excluídas e a estrutura RST é reestruturada, resultando na estrutura apresentada na Figura 20, na qual a ordem de importância das *EDUs* é permanecida.

[13] As *EDUs* candidatas ao sumário são todas aquelas consideradas mais importantes do que a *EDU* em foco.

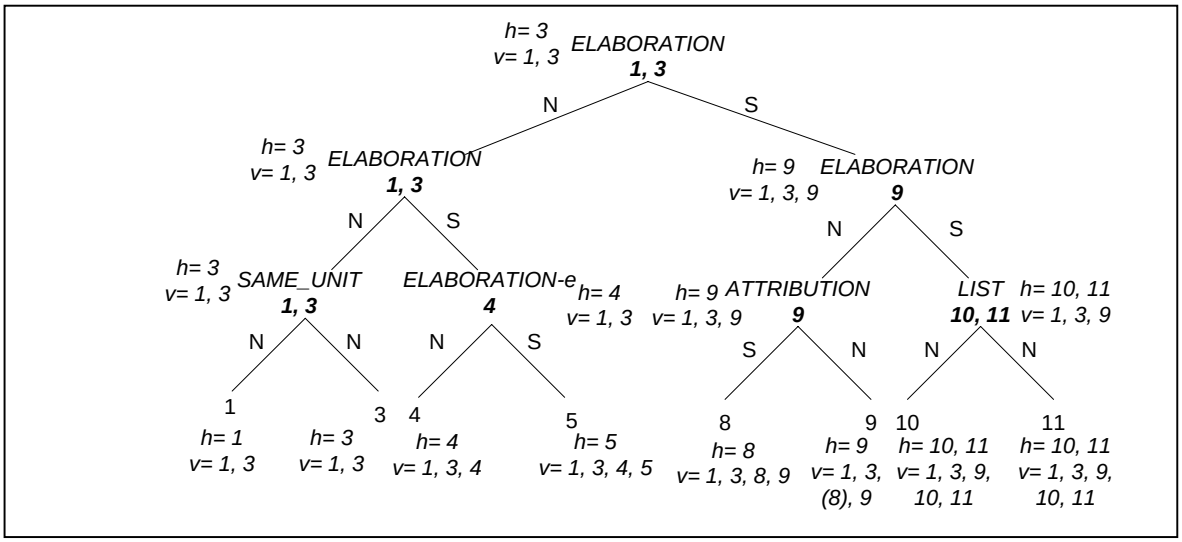

Figura 20: Estrutura RST intermediária 1

Porém, essa estrutura ainda não é a estrutura do sumário final, pois não se atingiu a taxa de compressão desejada. Assim, o processo se repete e, desta vez, apenas para a *EDU* 5 há uma heurística que se aplica. Essa heurística é a H11 e se refere à relação *ELABORATION-e*. A estrutura obtida após a aplicação dessa heurística é ilustrada na Figura 21.

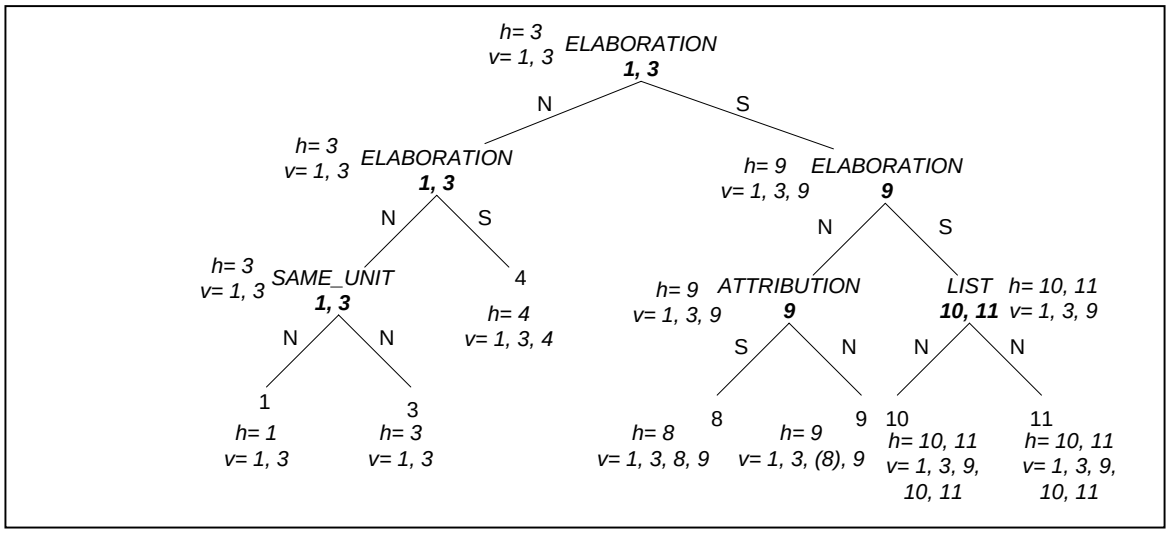

Figura 21: Estrutura RST intermediária 2

Como a taxa de compressão ainda não foi atingida, o processo se repete novamente e a heurística H10, que remete a relação *ELABORATION*, é aplicada, para a poda da *EDU* 4. A Figura 22 ilustra a estrutura RST resultante desse processo.

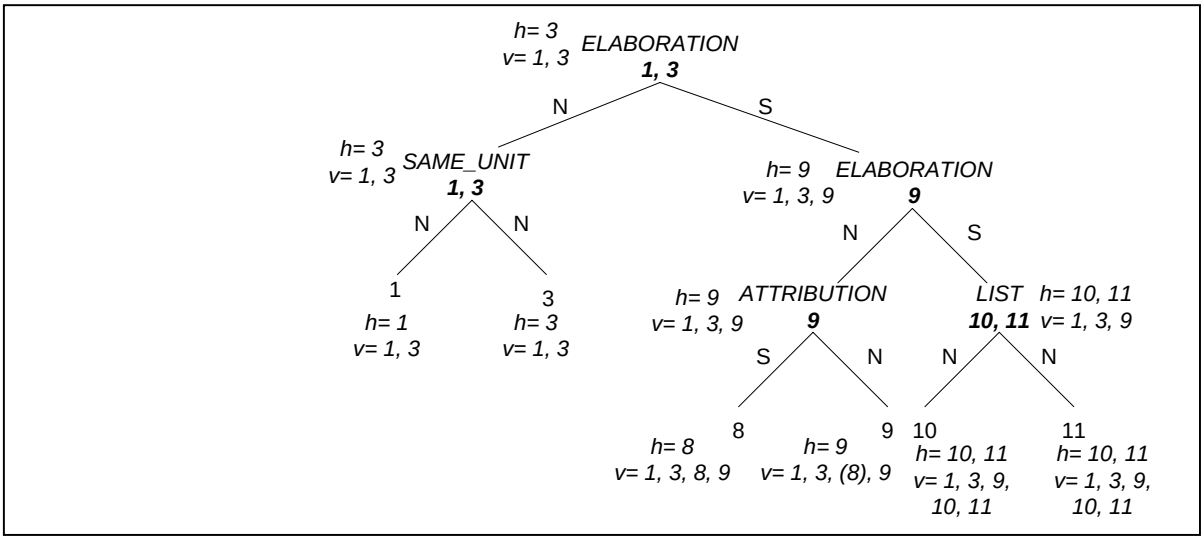

Figura 22: Estrutura RST intermediária 3

Uma vez que não há mais heurísticas a serem aplicadas e a taxa de compressão ainda não tenha sido atingida, as *EDUs* nucleares menos importantes são podadas, para que se atinja a taxa de compressão. Assim, as *EDUs* 10 e 11 também são podadas, resultando na estrutura RST do sumário ilustrada na Figura 23. Vale lembrar que a restrição de se incluir uma veia completa, quando uma de suas *EDUs* for incluída deve ser respeitada, mesmo que exceda um pouco a taxa de compressão. Por essa razão é que a estrutura RST do sumário final é composta por 4 *EDUs*.

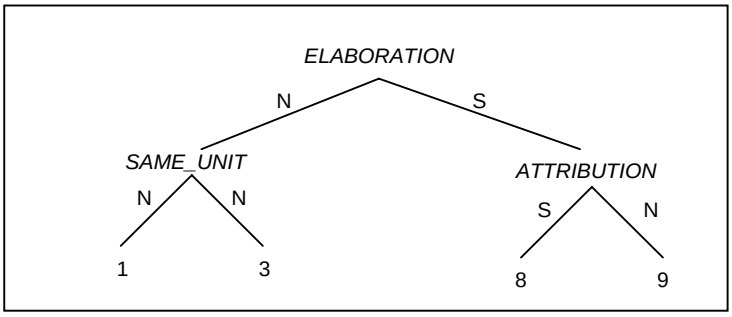

Figura 23: Estrutura RST do sumário

Uma possível realização lingüística dessa estrutura consiste em justapor as *EDUs* na ordem em que aparecem na estrutura do sumário. Assim, tem-se o sumário em língua natural apresentado na Figura 24. Como visto na seção anterior, a elaboração de um módulo real de realização lingüística não foi considerada neste trabalho.

> 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 24: Sumário subjacente à estrutura RST da Figura 20

Nota-se que o sumário apresentado na Figura 24 apresenta uma quebra de co-referência (em negrito na figura). Isso ocorre porque não há uma relação direta inter-relacionando a *EDU 8*, que contém o termo anafórico "**o programa de qualidade**", e a *EDU* 6, que contém o seu antecedente "**o Programa Sebrae de Qualidade Total**" (conforme ilustrado na Figura 19). Esses casos em que não ocorre um relacionamento direto entre a *EDU* anafórica e a *EDU* antecedente não são tratados pela *Veins Theory*. 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 7) e a coerência foram preservadas. O próximo capítulo apresenta uma avaliação mais detalhada do modelo proposto.

7. AVALIAÇÃO DO RHESUMARST

Com o objetivo de avaliar a qualidade do modelo de SA proposto, foram realizados dois experimentos. No primeiro, o RheSumaRST foi avaliado em relação à informatividade e em relação à coerência com os textos do corpus TeMário[14]. No segundo, somente a avaliação de coerência foi contemplada, usando textos do corpus Rhetalho. As seções 7.1 e 7.2 descrevem cada um deles, respectivamente.

7.1. Avaliação com o Corpus TeMário

O primeiro experimento teve dois propósitos principais: a) verificar se as heurísticas preservavam as informações mais relevantes do texto-fonte e b) verificar se as heurísticas garantiam a coerência dos sumários. No caso específico das cadeias de co-referências (CCRs), essa verificação consiste em investigar se os sumários apresentam quebra de CCRs. O corpus de teste, usado neste experimento, é composto de 10 textos também extraídos do TeMário (com um total de 5.277 palavras, aproximadamente 1 e ½ página cada texto). O número limitado de textos deveu-se, principalmente, ao tempo despendido na fase de preparação do corpus, que consistiu nas mesmas etapas da preparação do corpus usado para especificação das heurísticas (vide seção 6.1.2). As subseções 7.1.1 e 7.1.2 descrevem as avaliações de informatividade e de coerência, respectivamente.

7.1.1 Avaliação da Informatividade

Nesta etapa de avaliação, a ferramenta ROUGE (Lin (2004a); Lin (2004b))[15] foi utilizada. A ROUGE fornece várias medidas de cobertura para se obter automaticamente o grau de informatividade dos sumários automáticos. Tais medidas se baseiam na co-ocorrência de unidades de conteúdo como, por exemplo, n-gramas (ROUGE-N) e subseqüências de palavras em comum (ROUGE-L) entre os sumários automáticos e os

[14] Esses textos diferem daqueles usados no corpus de especificação das heurísticas.
[15] Essa ferramenta foi adotada na Document Understanding Conferences (DUC) de 2004 e 2005 (vide http://www-nlpir.nist.gov/projects/duc/data.html (último acesso: maio/2005)).

sumários construídos por humanos (aqui chamados de sumários ideais ou sumários de referência).

A medida ROUGE-N calcula a cobertura do sumário automático dividindo-se o número total de n-gramas do sumário de referência que co-ocorrem no sumário automático pelo número total de n-gramas do sumário de referência. Nessa medida, o "N" pode variar de 1 a 9, considerando-se n-gramas de vários tamanhos como, por exemplo, unigramas (ROUGE-1), bigramas (ROUGE-2) e assim por diante. A medida ROUGE-L é similar a ROUGE-N, exceto que ela considera a maior subseqüência de palavras em comum entre o(s) sumário(s) de referência(s) e o sumário automático, ao invés de considerar n-gramas[16]. Na avaliação do RHeSumaRST, somente as medidas de unigramas, bigramas, trigramas, quadrigramas e subseqüências mais longas de palavras em comum foram consideradas (e, portanto, ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4 e ROUGE-L). Essas medidas foram escolhidas porque são as mais usadas na DUC.

É válido dizer que a ROUGE permite a avaliação de um sumário automático usando um ou mais sumários de referência. No entanto, não há qualquer recomendação sobre o número de sumários de referência ideal. De acordo com Lin (2004a), à medida que se aumenta o número de sumários de referência, o índice de cobertura do sumário automático pode ser melhorado, pois a chance de o conteúdo do sumário automático estar presente nesses sumários é maior. O autor também argumenta que, independente do número de sumários de referência utilizado, a ferramenta consegue distinguir um sumário bom de um sumário ruim tão bem quanto um humano.

Para a avaliação do RHeSumaRST, pediu-se a cinco falantes nativos do português que construíssem um sumário de referência para cada um dos 10 textos do corpus, respeitando uma taxa de compressão de 70%, ou seja, cada sumário corresponderia a 30% do seu texto-fonte. Em outras palavras, para cada texto foram construídos cinco sumários de referência (na DUC' 2004, por exemplo, foram usados 4 sumários de referência). Os sumários automáticos também foram produzidos usando a mesma taxa de compressão. Logo, a cobertura foi calculada para cada sumário automático e seus cinco sumários de referência, aplicando-se a ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4 e ROUGE-L.

[16] Para mais detalhes sobre outras medidas vide (Lin, 2004a).

Adicionalmente, a cobertura do RHeSumaRST foi comparada com a de outros dois sistemas: o sistema proposto por Marcu (vide seção 5.1.2), aqui denominado Modelo de Saliência, e o sistema *Topline*. Este último é um baseline que poda todos os satélites de uma árvore RST mantendo apenas seus núcleos. Considera-se, neste caso, que ao preservar somente os núcleos de uma árvore RST (e, portanto, somente as informações mais importantes, segundo Mann and Thompson (1987)) é provável que se obtenham sumários altamente informativos. Devido a isso, ele é chamado de *Topline*.

Os dois sistemas escolhidos são similares ao RHeSumaRST na forma como classificam as *EDUs* para a poda, pois utilizam a função de saliência. Porém, a principal diferença, em relação ao RHeSumaRST, está no mecanismo de poda: como visto na seção 5.1.2, no Modelo de Saliência a poda é realizada obedecendo à classificação de importância das *EDUs* até que atinja a taxa de compressão desejada. Já no *Topline*, todos os satélites são excluídos após a classificação das *EDUs* e, em alguns casos, alguns núcleos menos salientes também são excluídos para satisfazer a taxa de compressão. Os resultados dessa avaliação são apresentados na Tabela 7 (somente as médias de cobertura obtidas em cada sistema são apresentadas).

Tabela 7: Graus de informatividade do RHeSumaRST considerando 5 sumários ideais

Sistema	**ROUGE-1**	**ROUGE-2**	**ROUGE-3**	**ROUGE-4**	**ROUGE-L**
Modelo de Saliência	0.55757	0.32286	0.25405	0.21768	0.53192
RHeSumaRST	0.57110	0.32640	0.25346	0.21921	0.54550
Topline	0.58424	0.33659	0.25960	0.21525	0.55663

Como se pode observar, o RHeSumaRST obteve os resultados mais próximos aos do *Topline*, quando ROUGE-1 e ROUGE-L foram usados. No entanto, ao aplicar as medidas ROUGE-2, ROUGE-3 e ROUGE-4 ambos RHeSumaRST e Modelo de Saliência tiveram performances muito similares. Esses resultados mostram que, embora o RheSumaRST mantenha também informações menos relevantes (isto é, satélites), para a preservação das veias, os sumários produzidos por ele podem ser tão informativos quanto os do *Topline*, que contém somente núcleos.

Com o propósito de verificar se a performance de cada sistema se mantém ao considerar um número menor de sumários de referência, cada sistema foi avaliado, também, com apenas três sumários de referência do conjunto de referência. Os resultados obtidos são mostrados na Tabela 8.

Tabela 8: Graus de informatividade do RHeSumaRST considerando 3 sumários ideais

Sistema	**ROUGE-1**	**ROUGE-2**	**ROUGE-3**	**ROUGE-4**	**ROUGE-L**
Modelo de Saliência	0.52877	0.30084	0.24144	0.21244	0.50245
RHeSumaRST	0.53738	0.29929	0.23549	0.20537	0.51080
Topline	0.56431	0.32326	0.25506	0.22386	0.53402

Embora o grau de cobertura em todos os sistemas tenha diminuído, ao se utilizar apenas três sumários de referência, o RHeSumaRST se manteve melhor que o Modelo de Saliência, quando aplicadas as ROUGE-1 e ROUGE-L. Porém, ao aplicar a ROUGE-2, ROUGE-3 e ROUGE-4, o Modelo de Saliência obteve resultados mais próximos ao do *Topline*.

Esses resultados confirmam a hipótese de que quanto maior o conjunto de sumários de referência mais alto são os índices de cobertura.

7.1.2 Avaliação da Coerência

Para verificar se as heurísticas preservam a coerência dos sumários automáticos, comparou-se (manualmente) cada sumário com seu correspondente texto-fonte anotado com as CCRs. Essa comparação teve como objetivo verificar ocorrências de quebras de co-referências nos sumários automáticos, isto é, casos em que apenas a anáfora aparece no sumário e, portanto, casos que introduzem quebras de coerência. Uma vez identificada uma possível quebra de co-referência no sumário, recorria-se ao seu correspondente texto-fonte a fim de verificar se a referência estava sendo introduzida pela primeira vez no texto ou se havia um antecedente para ela e, portanto, a confirmação da quebra da CCR. Para efeito de comparação, os sistemas usados na avaliação anterior também foram analisados. A Tabela 9 mostra o número de quebras de CCRs obtido por cada sistema e a sua representatividade

no corpus. É válido ressaltar que as anáforas diretas não foram computadas nas quebras de co-referências, pois, uma vez que apresentam anáfora e antecedente iguais, não introduzem quebra de CCRs nos sumários.

Tabela 9: Índice de quebras de CCRs do RHeSumaRST

Sistema	**# de CCRs dos sumários**	**# de quebras de CCRs**	**quebras de CCRs (%)**
Modelo de Saliência	81	12	15
Topline	89	7	8
RheSumaRST	93	5	5

Conforme a tabela, o RHeSumaRST apresentou o menor índice de quebras de CCRs nos sumários. Esse resultado é plausível, uma vez que nem o sistema *Topline* e nem o Modelo de Saliência propõem tratar explicitamente a preservação dos elos co-referenciais. Particularmente, o Modelo de Saliência não usa nenhum recurso para garantir a inclusão do antecedente de uma anáfora quando a mesma for inclusa no sumário, o que pode justificar seu pior desempenho. Já no caso do *Topline*, pode-se dizer que as estruturas RST espelham tão bem a organização dos textos, que a coerência dos sumários pode ser assegurada quase que independentemente dos satélites.

Se por um lado o RHeSumaRST apresentou o menor índice de quebra de co-referência, provando ser útil para tratar problemas de coerência introduzidos por quebras de CCRs, por outro lado, os resultados são bastante próximos aos obtidos pelo *Topline* e pelo Modelo de Saliência, para justificar todo o esforço necessário. Além do mais, o corpus usado nessa avaliação é muito pequeno para uma conclusão mais significativa. Devido a esses fatores, uma nova avaliação de coerência se fez necessária, como será descrito na seção a seguir.

7.2 Avaliação com o Corpus Rhetalho

Neste segundo experimento, somente a avaliação de coerência foi contemplada. O corpus adotado para essa avaliação foi o corpus Rhetalho[17]. Esse corpus é composto, atualmente, por uma coleção de 45 textos (25 do domínio da computação e 20 jornalísticos, com um total de 4.711 palavras (menor que ½ página)) anotados retoricamente segundo a RST. Esse corpus foi construído por dois especialistas em RST, que anotaram todos os textos, visando à produção de um corpus mais confiável. A ferramenta *RST Annotation Tool* foi usada para auxiliar a anotação. Para evitar discordâncias entre os anotadores, adotou-se um protocolo de anotação retórica (vide Apêndice B).

Adicionalmente, para esse experimento, cada texto do Rhetalho foi anotado com as CCRs (somente as descrições definidas). Os mesmos critérios adotados na anotação do corpus usado para a especificação das heurísticas (vide seção 6.1.2.2) foram seguidos aqui.

A tarefa de avaliação consistiu nas mesmas etapas da avaliação de coerência realizada com o corpus TeMário, conforme visto na seção anterior. Devido ao custo da avaliação manual, somente o sistema *Topline* foi usado, para comparação. As Tabelas 10 e 11 apresentam os resultados obtidos com os textos jornalísticos e científicos (isto é, do domínio da computação), respectivamente.

Tabela 10: Índice de quebras de CCRs do RHeSumaRST para textos jornalísticos

Sistema	**# de CCRs dos sumários**	**# de quebras de CCRs**	**quebras de CCRs (%)**
RheSumaRST	45	2	4
Topline	45	8	18

[17] Desenvolvido no contexto deste projeto e do projeto DiZer (Pardo et al., 2004), encontra-se disponível para download em: http://www.nilc.icmc.usp.br

Tabela 11: Índice de quebras de CCRs do RHeSumaRST para textos científicos

Sistema	# de CCRs dos sumários	# de quebras de CCRs	quebras de CCRs (%)
RheSumaRST	17	5	29
Topline	23	5	22

Embora o corpus de teste ainda seja pequeno para uma avaliação robusta, o RHeSumaRST se manteve melhor que o *Topline*, quando os sumários de textos jornalísticos foram avaliados. Entretanto, quando considerados os sumários de textos científicos, o *Topline* apresentou o menor índice de quebra de CCRs. Esses resultados podem indicar que as heurísticas propostas neste trabalho são dependentes de gênero, ou seja, um novo conjunto de heurísticas pode ser necessário para textos científicos. Porém, grandes corpora de textos devem ser avaliados para uma conclusão mais concreta.

7.3 Considerações Sobre os Experimentos

Este capítulo apresentou dois experimentos preliminares realizados para avaliar a viabilidade do modelo de SA proposto (vide Capítulo 6). O primeiro experimento contemplou a avaliação de informatividade e a avaliação de coerência. Já no segundo experimento, somente a avaliação de coerência foi contemplada.

Os resultados obtidos mostram que a maioria dos sumários gerados preserva as informações mais importantes do texto-fonte e são coerentes (no contexto deste trabalho, isto que dizer que eles não contêm quebras de cadeias de co-referências, mais, especificamente, de descrições definidas). Evidencia-se, assim, a utilidade do RHeSumaRST na produção de sumários de textos em língua natural, a partir de suas estruturas RST. Vale notar que, apesar de o RheSumaRST ter sido modelado com foco apenas nas descrições definidas, ele pode se aplicar a qualquer expressão lingüística de cadeias de co-referências, uma vez que seu mecanismo automático é independente de quaisquer fenômenos lingüísticos.

No capítulo a seguir apresentam-se as contribuições deste trabalho e também algumas possibilidades de trabalhos futuros.

8. CONTRIBUIÇÕES E TRABALHOS FUTUROS

Este trabalho trouxe várias contribuições para a área de sumarização automática, apontando também diversos trabalhos futuros, como apresentado nas próximas seções.

8.1 Contribuições

Destacam-se, nesta seção, vários tipos de contribuições obtidas com este trabalho. São elas:

- **Modelagem**

  - Definição de um novo modelo de sumarização automática, baseado no modelo de estruturação de discurso *RST* e no modelo de coerência global do discurso *Veins Theory.*
  - Especificação de um elenco de heurísticas de poda de estruturas RST de textos, baseadas em características específicas das relações retóricas da RST e em restrições de acessibilidade referencial da *Veins Theory*.

- **Protótipos**

  - Construção do protótipo RHeSumaRST, que permite a aplicação automática do modelo de sumarização proposto. Esse protótipo é importante para o desenvolvimento de testes e estudos, permitindo uma análise mais aprofundada do modelo na identificação de características a serem aprimoradas. Além disso, pode ser útil para a obtenção de textos condensados em um contexto de comunicação.
  - Construção de um sistema baseline (denominado *Topline*), visando a comparação com o RHeSumaRST.

- Implementação do Modelo de Saliência proposto por Marcu, para a comparação com o modelo proposto.

- **Corpora**

  - Construção de corpora anotados retoricamente
    - A produção de estruturas RST para 40 textos do corpus TeMário;
    - A construção do corpus Rhetalho, composto por 45 textos e suas subjacentes estruturas RST.

    O uso desses corpora na avaliação do RHeSumaRST resultou em outros três corpora:

    1) Corpus de sumários gerados com o RHeSumaRST e com os sistemas *Topline* e Modelo de Saliência, para 10 textos do TeMário, somando um total de 30 sumários;
    2) Corpus de sumários construídos por humanos para 10 textos do TeMário (mais especificamente, quatro sumários para cada texto, ou seja, 40 sumários no total);
    3) Corpus de sumários gerados com o RHeSumaRST e com o *Topline*, para os 45 textos do Rhetalho, somando um total de 90 sumários.
  - Produção de corpora anotados com as co-referências anafóricas definidas
    - Anotação de 40 textos do corpus TeMário;
    - Anotação de 45 textos do corpus Rhetalho.

- **Propostas Metodológicas**

  - Proposta de metodologia de desenvolvimento de um sumarizador automático de estruturas RST baseado na poda de informações irrelevantes.
  - Proposta de metodologia de avaliação de um sumarizador automático.

- **Outras**

- Identificação de algumas limitações da RST para a SA como, por exemplo, o fato de que ela não é capaz de indicar as informações irrelevantes cuja exclusão implique a quebra de coerência dos sumários (vide seção 1).
- Identificação das principais subestruturas de uma estrutura RST de um texto que indicam informações irrelevantes e cuja exclusão não implique a quebra de co-referência nos sumários, pelo menos, para o gênero e domínio de corpora utilizados neste trabalho. Por exemplo, as subestruturas formadas por *EDUs* satélites que são as folhas de uma estrutura RST.

8.2 Trabalhos Futuros

Apresentam-se, aqui, várias propostas de trabalhos futuros:

- **Extensão do RHeSumaRST**

Uma possibilidade de extensão refere-se à expansão do modelo de sumarização do RHeSumaRST, levando em consideração outras formas de sumarização como, por exemplo, a sumarização com base em foco específico, fornecido pelo usuário do sistema. Neste caso, seria possível explorar o modelo baseado na *Veins Theory* considerando veias específicas que se relacionam diretamente ao foco indicado pelo usuário, de modo similar ao realizado em Cristea et al. (2005).

- **Acoplamento de um Analisador Discursivo**

O protótipo RheSumaRST poderá ser acoplado ao sistema DiZer (Pardo 2005), que é um analisador discursivo automático de textos em português. Esse acoplamento consistirá em adaptar a estrutura RST de um texto obtida com o DiZer, em formato Prolog, para o formato de entrada do RHeSumaRST, ou seja, o formato SGML. Assim, teríamos um sistema com dois módulos, faltando somente o módulo de realização lingüística, para completar o sumarizador.

- **Construção de um Realizador Lingüístico**

Como visto anteriormente, o RHeSumaRST não contempla um módulo de realização lingüística e, devido a isso, a produção dos sumários em língua natural baseia-se somente na justaposição das unidades discursivas de sua estrutura RST. Uma possibilidade de torná-lo um sumarizador mais real é a construção de um realizador lingüístico capaz de realizar as estruturas RST dos sumários, gerando os sumários, propriamente dito, em língua natural.

- **Experimentos Visando a Escalabilidade do RHeSumaRST**

Visando tornar o RHeSumaRST um sistema mais robusto e real, outras avaliações devem ser realizadas, tanto do ponto de vista da coerência como do ponto de vista da informatividade. Em relação à primeira, o RHeSumaRST pode ser avaliado considerando-se outras expressões lingüísticas das cadeias de co-referências como, por exemplo, os pronomes pessoais, possessivos e demonstrativos. Neste caso, os mesmos corpora utilizados nos experimentos anteriores podem ser aproveitados, devendo ser anotados com as CCRs a serem consideradas. Outra possibilidade é a construção de novos corpora maiores, que possam ser utilizados também na condução de um experimento mais detalhado da informatividade, considerando-se o julgamento de humanos, para uma avaliação mais confiável do protótipo. Além dessas possibilidades, é possível, ainda, avaliar o comportamento do RHeSumaRST contemplando-se corpora de outros gêneros ou domínios como, por exemplo, científicos e jurídicos. Embora o modelo de sumarização do RHeSumaRST tenha sido proposto a partir da análise de corpus de textos escritos em português, corpora de textos escritos em outra língua natural também podem ser considerados, uma vez que o modelo é independente de língua.

No capítulo a seguir apresentam-se as considerações finais deste trabalho, bem como suas principais limitações.

9. CONSIDERAÇÕES FINAIS

Esta dissertação apresentou um modelo profundo de sumarização automática que incorpora um modelo cooperativo, adicionando à estruturação RST a proposta de relacionamento co-referencial da *Veins Theory*. Além dessas teorias, o modelo também agrega a função de saliência de unidades discursivas proposta por Marcu. O modelo proposto consiste em um elenco de heurísticas que são responsáveis por podar uma estrutura RST de um texto, para a produção do seu sumário. Tais heurísticas são guiadas por restrições de acessibilidade referencial da *Veins Theory*, visando à preservação da coerência dos sumários. O modelo foi implementado no protótipo RHeSumaRST, que independe de língua natural, uma vez que contempla apenas o nível de representação discursiva, factível pela estruturação da RST.

Três hipóteses foram assumidas neste trabalho: a primeira, de que é possível utilizar a estrutura RST de um texto para identificar suas informações irrelevantes; a segunda, de que é possível sumarizar um texto por meio da poda de informações irrelevantes da estrutura RST subjacente e a terceira, de que a poda dirigida por restrições de acessibilidade referencial da *Veins Theory* pode levar a sumários mais coerentes.

Experimentos preliminares foram realizados com o RheSumaRST para confirmá-las. Os resultados comprovaram sua validade: a maioria dos sumários produzidos preservam as informações mais relevantes do texto-fonte e são coerentes. No contexto, deste trabalho, sumários coerentes são aqueles que não apresentam quebra de cadeias de co-referências, mais, especificamente, de co-referências anafóricas definidas.

Embora o foco deste trabalho seja a produção de sumários a partir de estruturas RST de textos, a ausência de um analisador discursivo, que processe um texto em língua natural produzindo sua estrutura RST, e de um realizador lingüístico, que realize superficialmente a estrutura RST do sumário, produzindo o sumário em língua natural, não faz do RHeSumaRST um sumarizador real. Isto se deve ao fato de que, para o seu funcionamento, faz-se necessária a intervenção humana, principalmente, na fase de construção das estruturas RST. Entretanto, o RHeSumaRST pode ser muito útil em um futuro próximo ao acoplá-lo ao sistema DiZer (Pardo et al. 2004; Pardo 2005), um analisador discursivo automático de textos escritos em português do Brasil. Dessa forma, um sumarizador

automático de dois módulos estará disponível, faltando apenas um módulo de realização lingüística para completar o sumarizador. Muito embora seja uma alternativa de longo prazo, já há no NILC um trabalho em desenvolvimento nessa linha de pesquisa. Mais especificamente, o trabalho de Pelizzoni and Nunes (2005) trata do desenvolvimento de um realizador lingüístico para a geração de textos em português.

Na seção seguinte destacam-se as principais limitações deste trabalho.

9.1 Limitações deste Trabalho

Este trabalho apresenta algumas limitações que se tornaram claras no decorrer dessa dissertação. São elas:

- O RHeSumaRST não é um sumarizador de fato, pois seu modelo de sumarização consiste, simplesmente, na poda de informações irrelevantes de uma estrutura RST de um texto, não havendo um processo real de condensação de conteúdo.
- Complementarmente, o modelo proposto não contempla um analisador discursivo para a construção da estrutura RST do texto a ser sumarizado. O usuário do RHeSumaRST deve ser especialista em RST para poder construir a estrutura retórica do texto com a ferramenta *RST Annotation Tool*, a qual servirá de entrada para o sistema.

REFERÊNCIAS BIBLIOGRÁFICAS

Bick, Eckhard. (2000). *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD Thesis, Arhus University, Arhus.

Carlson, L. and Marcu, D. (2001). *Discourse Tagging Reference Manual.* Technical Report ISI-TR-545, University of Southern, California.

Coelho, J.C.B. (2004). *Cadeias de co-referência aplicadas à sumarização automática.* Mostra de Iniciação Científica - MIC'2004. UNISINOS, São Leopoldo - RS.

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.

Cristea, D.; Postolache, O.; Pistol, I. (2005). Summarization Through Discourse Structure. *In the Proceedings of the 6th International Conference on Computational Linguistics and Intelligence Text Processing – CICLing 2005,* Mexico.

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.

Lin, C. (2004a). ROUGE: a Package for Automatic Evaluation of Summaries. *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004),* Barcelona, Spain.

Lin, C. (2004b). Looking for a Few Good Metrics: Automatic Summarization Evaluation - How Many Samples Are Enough?. *In Proceedings of the NTCIR Workshop 4*, Tokyo, Japan.

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

Marcu, D. (1997b). From Discourse Structures to Text Summaries. *In the Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization,* pp. 82-88. Madrid, Spain.

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.

Milner, J. C. (2003). Reflexões sobre a referência e a correferência. *In* M.M. Cavalcante, B.B. Rodrigues, A. Ciulla. (eds.)*, Referenciação*. Editora Contexto.

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.

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

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. *In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken*, pp. 210-218 (Lecture Notes in Artificial Intelligence 2721). Springer-Verlag, Germany.

Pardo, T.A.S; Nunes, M.G.V.; Rino, L.H.M. (2004). *DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese.* XVII Brazilian Symposium on Artificial Intelligence - SBIA'04, São Luís - Maranhão.

Pardo, T.A.S. (2005). Métodos para Análise Discursiva Automática. Tese de Doutorado. ICMC/USP, São Carlos-SP.

Pelizzoni, J.M. and Nunes, M.G.V. (2005). Reconciling Parameterization, Configurability and Optimality in Natural Language Generation via Multiparadigm Programming. In *Proceedings of Sixth International Conference on Intelligent Text Processing and Computational Linguistics - CICLing 2005*, Mexico City, Mexico.

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.

Rino, L.H.M. e Pardo, T.A.S. (2003). A Sumarização Automática de Textos: Principais Características e Metodologias. *Anais do XXIII Congresso da Sociedade Brasileira de Computação*, Vol. VIII: III Jornada de Minicursos de Inteligência Artificial (III MCIA), pp. 203-245. Campinas-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.

Seno, E.R.M.; Rino, L.H.M. (2005a). *Heurísticas de Sumarização de Estruturas RST*. Série de Relatórios Técnicos: NILC-TR-05-04, ICMC/USP, São Carlos, Brasil.

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

Sparck-Jones, K. (1993b). What might be in a summary? In G. Knorz; J. Krause and C. Womser-Hacker (eds.), *Information Retrieval 93*, pp. 9-26. Universitatsverlag Konstanz.

Vieira, R. and Salmon-Alt, S. (2002). Nominal Expression in Multilingual Corpora: Definite and Demonstratives. In *the Proceedings of the Language Resources and Evaluation Conference - LREC 2002*, Las Palmas.

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.

Apêndice A – Definição das Relações Retóricas

Neste apêndice, são apresentadas as definições das relações retóricas utilizadas neste trabalho. A definição de cada relação consiste de quatro tipos de informações que o analista de um texto deve considerar, para determinar como as *EDUs* se inter-relacionam (Mann and Thompson, 1987). 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.

As Figuras A.1-A.25 apresentam as definições de cada relação. As relações retóricas que também podem ser encaixadas (representadas por (-e)) são definidas uma única vez, pois não há diferença em relação à definição daquelas não encaixadas.

Nome da relacão: ATTRIBUTION
Restrições sobre N: N apresenta uma expressão, fala ou pensamento de alguém ou algo **Restrições sobre S**: S apresenta alguém ou algo que produz N **Restrições sobre N+S**: S e N indicam, respectivamente, a fonte de uma mensagem e a mensagem **Efeito**: o leitor é informado sobre a mensagem e sobre quem ou o que a produziu

Figura A.1 – Definição da relação ATTRIBUTION

Nome da relação: CAUSE (-e)
Restrições sobre N: apresenta a causa de uma situação **Restrições sobre S**: apresenta o resultado de uma situação **Restrições sobre N+S**: N apresenta uma situação que é a causa da situação apresentada em S; sem N, o leitor poderia não reconhecer o que causou a situação apresentada em S; N é mais central para a satisfação do objetivo do escritor do que S **Efeito**: o leitor reconhece a situação apresentada em N como a causa da ação apresentada em S

Figura A.2 – Definição da relação CAUSE

Nome da relação: CIRCUMSTANCE (-e)
Restrições sobre N: não há **Restrições sobre S**: apresenta uma situação (realizável) **Restrições sobre N+S**: S provê uma situação na qual o leitor pode interpretar N **Efeito**: o leitor reconhece que S provê uma situação na qual N deve ser interpretado

Figura A.3 – Definição da relação CIRCUMSTANCE

Nome da relação: COMPARISON (-e)
Restrições sobre N: apresenta uma característica de algo ou alguém **Restrições sobre S**: apresenta uma característica de algo ou alguém comparável com o que é apresentado em N **Restrições sobre N+S**: as características de S e N estão em comparação **Efeito**: o leitor reconhece que S é comparado a N em relação a certas características

Figura A.4 – Definição da relação COMPARISON

Nome da relação: CONCESSION
Restrições sobre N: o escritor julga N válido **Restrições sobre S**: o escritor não afirma que S pode não ser válido **Restrições sobre N+S**: o escritor mostra uma incompatibilidade aparente ou em potencial entre N e S; o reconhecimento da compatibilidade entre N e S melhora a aceitação de N pelo leitor **Efeito**: o leitor aceita melhor N

Figura A.5 – Definição da relação CONCESSION

Nome da relação: CONDITION
Restrições sobre N: não há **Restrições sobre S**: S apresenta uma situação hipotética, futura ou não realizada **Restrições sobre N+S**: a realização de N depende da realização de S **Efeito**: o leitor reconhece como a realização de N depende da realização de S

Figura A.6 – Definição da relação CONDITION

Nome da relação: CONTRAST
Restrições sobre os Ns: não mais do que dois Ns; as situações nos Ns são (a) compreendidas como similares em vários aspectos, (b) compreendidas como diferentes em vários aspectos e (c) comparadas em relação a uma ou mais dessas diferenças **Efeito**: o leitor reconhece as similaridades e diferenças resultantes da comparação sendo feita

Figura A.7 – Definição da relação CONTRAST

Nome da relação: ELABORATION (-e)
Restrições sobre N: não há **Restrições sobre S**: não há **Restrições sobre N+S**: S apresenta detalhes adicionais sobre a situação ou algum elemento de N **Efeito**: o leitor reconhece que S apresenta detalhes adicionais sobre N

Figura A.8 – Definição da relação ELABORATION

Nome da relação: EVIDENCE
Restrições sobre N: o leitor poderia não acreditar em N de forma satisfatória para o escritor **Restrições sobre S**: o leitor acredita em S ou o achará válido **Restrições sobre N+S**: a compreensão de S pelo leitor aumenta sua convicção em N **Efeito**: a convicção do leitor em N aumenta

Figura A.9 – Definição da relação EVIDENCE

Nome da relação: EXAMPLE
Restrições sobre N: não há **Restrições sobre S**: apresenta um exemplo de algo ou de uma situação **Restrições sobre N+S**: S apresenta um exemplo de algo ou de uma situação apresentada em N **Efeito**: o leitor reconhece S como um exemplo de algo ou de uma situação apresentada em N

Figura A.10 – Definição da relação EXAMPLE

Nome da relação: EXPLANATION-ARGUMENTATIVE (-e)
Restrições sobre N: apresenta um evento ou situação **Restrições sobre S**: não há **Restrições sobre N+S**: S apresenta uma explicação para o evento ou situação apresentado em N **Efeito**: o leitor reconhece que S fornece uma explicação para o evento ou situação apresentado em N

Figura A.11 – Definição da relação EXPLANATION-ARGUMENTATIVE

Nome da relação: INTERPRETATION
Restrições sobre N: não há **Restrições sobre S**: não há **Restrições sobre N+S**: S apresenta um conjunto de idéias que não é expresso em N propriamente, mas derivado deste **Efeito**: o leitor reconhece que S apresenta um conjunto de idéias que não é propriamente expresso no conhecimento fornecido por N

Figura A.12 – Definição da relação INTERPRETATION

Nome da relação: JOINT
Restrições sobre os Ns: não há **Efeito**: não há

Figura A.13 – Definição da relação JOINT

Nome da relação: JUSTIFY (-e)
Restrições sobre N: não há **Restrições sobre S**: não há **Restrições sobre N+S**: a compreensão de S pelo leitor aumenta sua prontidão para aceitar o direito do escritor de apresentar N **Efeito**: a prontidão do leitor para aceitar o direito do escritor de apresentar N aumenta

Figura A.14 – Definição da relação JUSTIFY

Nome da relação: LIST
Restrições sobre os Ns: itens comparáveis apresentados nos Ns **Efeito**: o leitor reconhece como comparáveis os itens apresentados

Figura A.15 – Definição da relação LIST

Nome da relação: MEANS (-e)
Restrições sobre N: uma atividade **Restrições sobre S**: não há **Restrições sobre N+S**: S apresenta um meio, método ou instrumento que faz com que a atividade em N seja realizada **Efeito**: o leitor reconhece que o meio, método ou instrumento em S faz com que a atividade em N seja realizada

Figura A.16 – Definição da relação MEANS

Nome da relação: PARENTHETICAL
Restrições sobre N: não há **Restrições sobre S**: apresenta informação extra relacionada a N que não está expressa no fluxo principal do texto **Restrições sobre N+S**: S apresenta informação extra relacionada a N, complementando N; S não pertence ao fluxo principal do texto **Efeito**: o leitor reconhece que S apresenta informação extra relacionada a N, complementando N

Figura A.17 – Definição da relação PARENTHETICAL

Nome da relação: PURPOSE (-e)
Restrições sobre N: apresenta uma atividade **Restrições sobre S**: apresenta uma situação não realizada **Restrições sobre N+S**: S apresenta uma situação que será realizada através da atividade apresentada em N

Figura A.18 – Definição da relação PURPOSE

Nome da relacão: REASON (-e)
Restrições sobre N: apresenta uma situação **Restrições sobre S**: apresenta a razão de uma situação **Restrições sobre N+S**: S é a razão para a situação apresentada em N **Efeito**: o leitor reconhece que S é a razão para a situação apresentada em N

Figura A.19 – Definição da relação REASON

Nome da relação: RESULT
Restrições sobre N: apresenta o resultado de uma situação **Restrições sobre S**: apresenta a causa de uma situação **Restrições sobre N+S**: N apresenta o resultado de uma situação causada pela situação apresentada em S; sem S, o leitor poderia não reconhecer o que causou a situação apresentada em N; N é mais central para a satisfação do objetivo do escritor do que S **Efeito**: o leitor reconhece a situação apresentada em N como um resultado da situação causada por S

Figura A.20 – Definição da relação RESULT

Nome da relação: SAME-UNIT
Restrições sobre os Ns: os Ns apresentam informações que, juntas, constituem uma única proposição **Efeito**: o leitor reconhece que as informações apresentadas constituem uma única proposição; separadas, não fazem sentido

Figura A.21 – Definição da relação SAME-UNIT

Nome da relação: SEQUENCE
Restrições sobre os Ns: as situações apresentadas nos Ns são realizadas em seqüência **Efeito**: o leitor reconhece a sucessão temporal dos eventos apresentados

Figura A.22 – Definição da relação SEQUENCE

Nome da relação: SUMMARY (-e)
Restrições sobre N: não há **Restrições sobre S**: não há **Restrições sobre N+S**: S apresenta um resumo do conteúdo de N **Efeito**: o leitor reconhece S como um resumo do conteúdo de N

Figura A.23 – Definição da relação SUMMARY

Nome da relação: TEMPORAL-AFTER
Restrições sobre N: não há **Restrições sobre S**: não há **Restrições sobre N+S**: N apresenta uma situação que ocorre depois da situação apresentada em S; N é mais central para a satisfação do objetivo do escritor do que S **Efeito**: o leitor reconhece que N apresenta uma situação que ocorre depois da situação apresentada em S

Figura A.24 – Definição da relação TEMPORAL-AFTER

Nome da relação: TEMPORAL-SAME-TIME
Restrições sobre N: não há **Restrições sobre S**: não há **Restrições sobre N+S**: N apresenta uma situação que ocorre em paralelo a situação apresentada em S; N é mais central para a satisfação do objetivo do escritor do que S **Efeito**: o leitor reconhece que N apresenta uma situação que ocorre em paralelo a situação apresentada em S

Figura A.25 – Definição da relação TEMPORAL-SAME-TIME

APÊNDICE B – PROTOCOLO DE ANOTAÇÃO RETÓRICA

Este apêndice apresenta o protocolo de análise retórica usado na construção do corpus Rhetalho.

Estratégia de Anotação

A anotação retórica deve ser linear, da esquerda para a direita. Primeiramente, devem-se relacionar todas as orações (*EDUs*) presentes em uma sentença; depois, todas as sentenças de um parágrafo; por fim, todos os parágrafos do texto devem ser relacionados, formando uma única estrutura retórica. Somente estruturas binárias são permitidas.

Critério de Segmentação

Para a segmentação dos textos, as regras propostas por Carlson and Marcu (2001) devem ser seguidas. Embora essas regras tenham sidas definidas para a língua inglesa, elas são genéricas o bastante para serem utilizadas na língua portuguesa.

Se houver discordância entre os anotadores em algum ponto da segmentação, deve-se adotar uma segmentação mais genérica e compreensiva.

Determinação de Relações Retóricas (incluindo a determinação de núcleos e satélites)

Relações estendidas devem ser usadas para relacionar subestruturas RST. Neste caso, deve-se seguir o critério de composicionalidade de Marcu (1997a). Somente as relações retóricas do conjunto pré-selecionado devem ser consideradas.

Se houver discordância entre os anotadores ao determinar a relação entre pares de segmentos discursivos, uma relação mais genérica deve ser escolhida. Porém, se ambas as relações forem igualmente plausíveis, um terceiro especialista em RST deve ser consultado para apontar a relação mais apropriada.