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

Análise Discursiva para a Sumarização Automática de Textos em Português

Eloize Rossi Marques Seno
Lucia Helena Machado Rino

NILC-TR-04-06

Agosto, 2004

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

Resumo

Este relatório apresenta os resultados da análise discursiva de um corpus de textos jornalísticos em português do Brasil com base na *Rhetorical Structure Theory – RST* (Mann and Thompson, 1987). A análise de discurso foi realizada como primeiro passo para elaboração de uma proposta de modelo computacional para a Sumarização Automática, no âmbito de um projeto de mestrado.

.

Índice

1. Introdução

O acúmulo excessivo de textos de diversas naturezas de que dispomos hoje e o tempo cada vez mais reduzido que as pessoas têm para absorver o máximo de informações em um curto espaço de tempo tem impulsionado cada vez mais as pesquisas na área de sumarização automática de textos.

Quando se fala em sumarização de textos, é necessário dizer o que se entende por sumários. Em poucas palavras pode-se dizer que um sumário é um texto condensado de uma fonte, o qual preserva o seu conteúdo mais relevante sem perder o significado original pretendido (Rino e Pardo, 2003). Manchetes de jornais, sinopses de novelas, artigos de revistas, *abstracts* de livros e teses são exemplos de sumários. Cada um desses tipos tem suas características particulares, assim como conteúdos e correspondência de vários teores com sua fonte. Por exemplo, há sumários que são informativos e contemplam as informações mais relevantes da sua fonte, dispensando, assim, sua leitura. Outros sumários são indicativos e servem apenas para indicar qual o conteúdo do texto-fonte a que se referem, sendo de grande utilidade, por exemplo, quando se quer fazer uma busca por documentos referentes a um determinado assunto. Outros assumem função de avaliadores, apresentando apenas uma avaliação dos conteúdos de suas fontes, são chamados sumários críticos (Mani, 2001).

A sumarização automática de textos vem sendo explorada desde o final da década de 50, quando começaram a surgir alguns métodos estatísticos para extrair as informações mais relevantes de um texto. Um dos trabalhos mais importantes daquela época foi o trabalho de Luhn (1958), que propôs um método baseado na freqüência de palavras-chave para escolher as sentenças mais importantes de um texto para compor o sumário. As pesquisas continuaram nas décadas seguintes, trazendo vários avanços para a área, sob a ótica de duas abordagens principais para a sumarização: a superficial e a profunda. A abordagem superficial utiliza métodos estatísticos e empíricos, enquanto a abordagem profunda utiliza teorias formais e modelos lingüísticos[1]. Um exemplo clássico de abordagem profunda se baseia na *RST – Rhetorical Structure Theory* (Mann and Thompson, 1987).

A RST é uma das teorias de discurso mais populares para a geração de língua natural que tem sido muito explorada por vários autores para a sumarização automática de textos. Por exemplo, Ono et al. (1994) desenvolveram um sistema que se baseia na estrutura RST de textos escritos em japonês para computar o grau de importância de cada unidade do discurso, considerando que as unidades mais importantes apresentam informações mais relevantes do texto, sendo adequadas, portanto, para compor o sumário. Marcu (1997a) utiliza as estruturas retóricas, extraídas automaticamente, de textos escritos em inglês, para computar a saliência das unidades discursivas a fim de formar os sumários. Diferentemente do trabalho de Ono et al., o sistema proposto por Marcu baseia-se na profundidade das unidades discursivas na estrutura RST. O'Donnel (1997a) propôs um sistema para a sumarização de documentos on-line também baseado nas proposições mais relevantes das estruturas retóricas dos textos. Entretanto, esse sistema não constrói automaticamente a estrutura retórica do texto a ser sumarizado, como é o caso dos sistemas de Ono et al. e Marcu, requerendo que a análise RST dos textos seja feita previamente.

Segundo esses autores, a RST fornece uma ordem natural de importância das unidades discursivas e, portanto, pode ser usada de modo eficiente na SA por permitir identificar as informações mais relevantes em um texto para compor o seu sumário.

[1] Para mais detalhes sobre as duas abordagens veja (Rino e Pardo, 2003).

Tendo em vista essa motivação, este relatório apresenta uma análise discursiva, segundo a RST, de um corpus de textos jornalísticos escritos em português para a proposta de um modelo computacional para a Sumarização Automática (SA).

A próxima seção apresenta uma breve introdução da RST. Na seção 3, apresenta-se a ferramenta de estruturação retórica (*RST Annotation Tool*) usada para analisar o corpus para a seguir, na seção 4, apresentar a análise discursiva. Por fim, a seção 5 apresenta as considerações finais.

2. Rhetorical Structure Theory

A *Rhetorical Structure Theory (RST)* foi desenvolvida na década de 80 por William C. Mann e Sandra Thompson com o objetivo de prover uma estrutura para a interpretação de línguas naturais. Atualmente, é a teoria de discurso mais usada na SA.

A RST parte do 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 pretendia 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 e são responsáveis por atribuir coerência a um texto.

Segundo a RST, as relações retóricas inter-relacionam *EDUs* que são expressas por segmentos adjacentes em um texto. A cada *EDU* é atribuído um papel de núcleo ou satélite. O núcleo, ou unidade nuclear, expressa a informação principal, sendo, portanto, mais relevante do que o satélite. O satélite apresenta informação adicional, a qual exerce influência na interpretação do leitor sobre a informação apresentada no núcleo. Assim, núcleos, na maioria das vezes, são compreensíveis independentemente dos satélites, mas não vice-versa.

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, como é o caso das relações multinucleares.

Segundo Mann e Thompson, o conjunto de relações retóricas da RST é capaz de representar todas as possíveis relações de significado entre os segmentos discursivos de uma grande gama de textos. Esse conjunto de relações é mostrado na Tabela 1[2].

[2] Mantidas com a nomenclatura original, suas definições podem ser recuperadas da obra de referência (Mann and Thompson, 1987).

Tabela 1: Conjunto de relações retóricas

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

Considere, por exemplo, as Figura 1 e 2, as quais ilustram relações mononuclear e multinuclear, respectivamente. No texto da Figura 1 (com as unidades elementares numeradas, para referência), a unidade elementar 1 é o satélite (S) e a unidade elementar 2 é o núcleo (N) da relação retórica *PURPOSE*, cujo satélite apresenta uma situação que será realizada através da atividade apresentada no núcleo. Em outras palavras pode-se dizer que uma das possíveis interpretações dessa estrutura seria N a fim de realizar S, como ilustra o texto correspondente. No texto da Figura 2 há uma relação retórica *SEQUENCE* indicando a seqüência de eventos entre as unidades elementares 1, 2 e 3, sendo que todas elas possuem o mesmo grau de importância. Como mostra a Tabela 1, o conjunto de relações da RST inclui apenas três relações multinucleares.

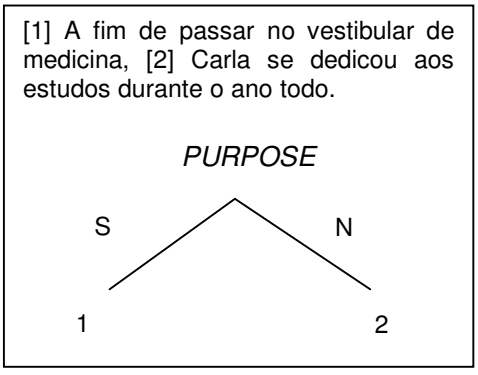

Figura 1: Exemplo de relação mononuclear

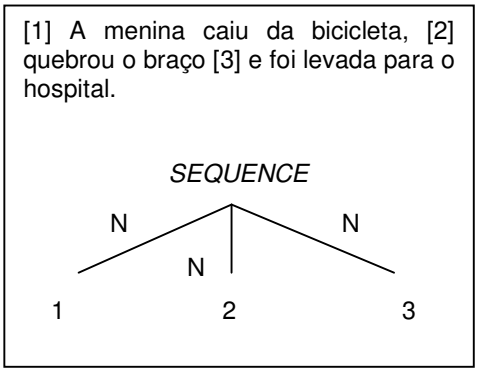

Figura 2: Exemplo de relação multinuclear

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

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

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

Tabela 2: Exemplos de definição de relações retóricas

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

Um texto pode ter sua estrutura discursiva representada por uma árvore retórica, aplicando-se relações individuais a pares de segmentos que variam, em tamanho, de uma simples oração até segmentos mais complexos, estruturados como subárvores inteiras. Assim, a estruturação retórica é recorrente. Uma árvore retórica é uma árvore

cujos nós folha correspondem às *EDUs* e cujos nós internos representam relações retóricas, como mostram as Figuras 1 e 2.

Ao desenvolver o primeiro analisador retórico automático para o inglês, Marcu (1997a, 2000) utilizou corpora de textos cuja análise indicava a existência de relações de significado não previstas no conjunto de relações da RST. Assim, ele modificou o conjunto de relações da RST introduzindo relações retóricas encaixadas (*embedded*, no inglês), que relacionam segmentos encaixados no discurso. Para diferenciar as relações retóricas encaixadas das não encaixadas, neste relatório, acrescenta-se um "-e" no final do nome dessas relações.

Como exemplo de segmento encaixado, considere a sentença "O livro de matemática que comprei em São Paulo é muito didático.", na qual a oração relativa "que comprei em São Paulo" é o segmento encaixado, o qual especifica o livro referenciado na sentença. Para representar esses casos, em que há segmentos encaixados constituindo uma única unidade elementar, faz-se uso da relação multinuclear *SAME-UNIT* proposta por Marcu (1997a). A Figura 3 mostra a estrutura retórica desta sentença.

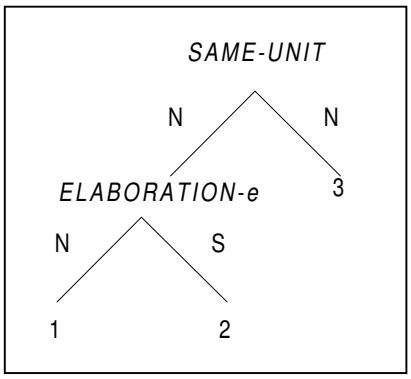

Figura 3: Estrutura retórica com segmento encaixado

A análise de discurso do corpus foi realizada com base na RST com o auxílio de uma ferramenta de estruturação retórica: a *RST Annotation Tool*. A análise também contou com o auxilio da ferramenta RhetDB[3], 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. A próxima seção apresenta brevemente a *RST Annotation Tool* para, então, apresentar a análise discursiva do corpus, na seção 4.

3. A Ferramenta de Estruturação Retórica

A *RST Annotation Tool*[4] é uma variação da ferramenta *RSTTool* (O'Donnell, 1997b). Ela 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 sobre técnicas de análise de discurso.

[3] Desenvolvida por Thiago A. S. Pardo em seu doutorado (Pardo, 2003).
[4] Disponível em: http://www.isi.edu/~marcu/discourse/AnnotationSoftware.html (acessada em dez/2003)

Os recursos disponíveis para a análise completa de um texto incluem dispositivos para segmentação, ou seja, especificação de uma *EDU*[5], identificação de seu papel retórico (isto é, se a *EDU* é N ou S) e identificação de seu inter-relacionamento (isto é, busca de uma relação retórica que expresse o relacionamento entre pares de *EDUs*). A ferramenta também permite alterar o conjunto de relações retóricas a ser utilizado, manipular estruturas retóricas já prontas como, por exemplo, desfazer operações e alterar relações, e ainda salvar as análises em arquivos com formatos LISP ou SGML. Outra vantagem desta ferramenta é que ela permite diversas estratégias de análise retórica como, por exemplo, análise incremental, conforme discutido em (Carlson and Marcu, 2001). A estratégia de análise usada para anotar o corpus será discutida na seção 4.

A Figura 4 mostra a interface da *RST Annotation Tool*. A parte inferior da janela apresenta, ao usuário, o texto a ser estruturado e a parte superior apresenta sua árvore RST em construção. O usuário delimita cada *EDU* no próprio texto simplesmente "clicando" no ponto que delimita seu fim. Como resultado, essa *EDU* é numerada, como se pode ver na parte inferior da janela. As folhas da árvore RST em construção são, assim, identificadas por seus números, não-ambíguos para um mesmo texto.

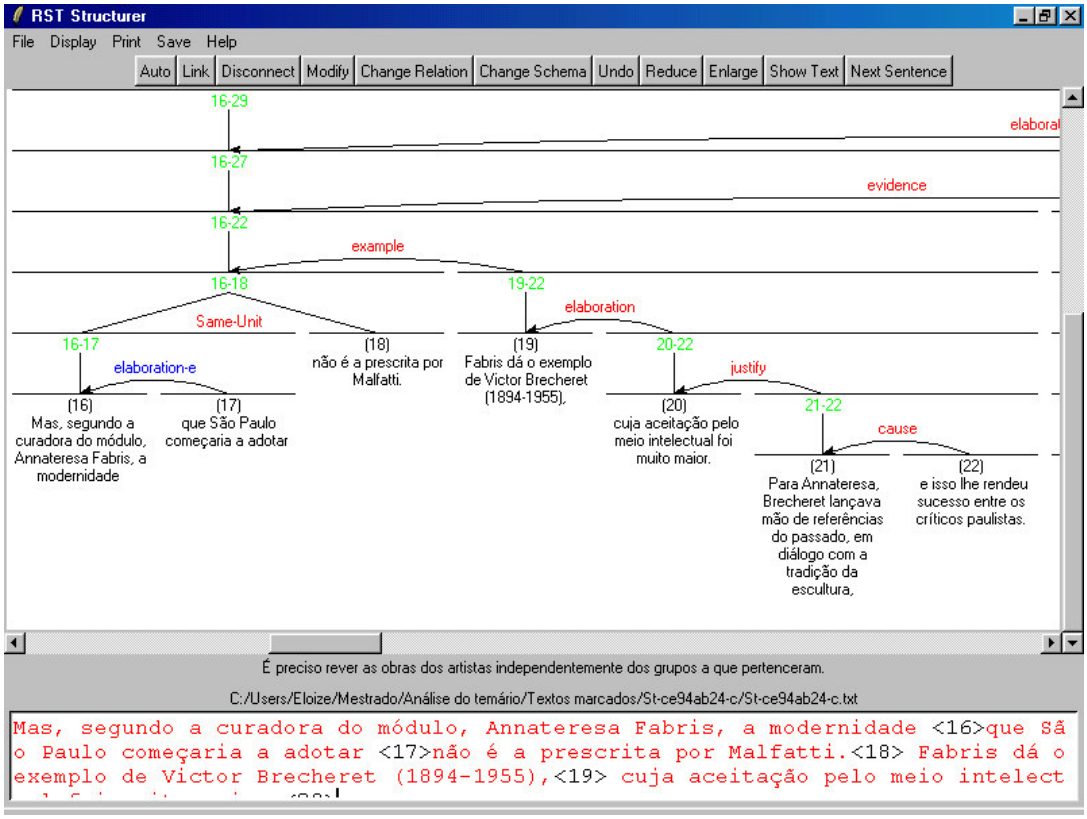

Figura 4: Interface da *RST Annotation Tool*

[5] Lembrando que cada *EDU* corresponde a uma unidade proposicional, isto é, uma unidade mínima de significado. Na análise RST, em geral, toma-se a representação superficial de unidades textuais como *EDUs*.

4. Análise Discursiva do Corpus

Elaborou-se a análise de discurso de 30 textos do gênero jornalístico, os quais foram extraídos do corpus TeMário (Pardo e Rino, 2003). Os textos variam de 1 a 2 ½ páginas. Foram escolhidos textos jornalísticos pelo fato de apresentarem uma linguagem bastante abrangente e de fácil compreensão, facilitando a interpretação e, portanto, a análise discursiva.

Para a análise foram utilizados o conjunto de relações retóricas da RST e também algumas relações propostas por (Marcu 1997a; Carlson and Marcu 2001).

A seguir são apresentados a forma como os textos foram segmentados para análise (subseção 4.1), a estratégia de análise adotada (subseção 4.2), um exemplo de análise discursiva (subseção 4.3), o conjunto de relações retóricas utilizado (subseção 4.4), alguns dados extraídos do corpus após a análise (subseção 4.5) e algumas dificuldades encontradas durante o processo de análise (subseção 4.6).

4.1 Segmentação Textual

Na RST, um texto pode ser segmentado com diversas granularidades, por exemplo, em parágrafos, frases, sentenças, orações, dentre outros. Para segmentar os textos do corpus adotaram-se orações como *EDUs*, seguindo basicamente a proposta de (Carlson and Marcu, 2001). Embora as regras de segmentação de discurso dessa proposta sejam fortemente dependentes da sintaxe, elas têm-se 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 também foram usados para ajudar a determinar as *EDUs*. Alguns exemplos dessas regras são: (Carlson and Marcu, 2001, p. 26, 27, 39, 40, 41)[6].

- Orações sinalizadas por marcadores discursivos fortes como, por exemplo, *Porque, Apesar de, Conforme, Segundo, Em conseqüência de, entre outros,* são consideradas *EDUs*;
- Orações principais são consideradas *EDUs*;
- Orações subordinadas com marcadores discursivos são consideradas *EDUs*;
- Orações complementares não são consideradas *EDUs*, exceto complemento de verbos de atribuição. Por exemplo: *[1] A companhia disse [2] que 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 (*embedded*, no inglês).

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, um texto pode conter muitas *EDUs* encaixadas, as quais são responsáveis por manter sua coerência.

[6] Tradução nossa.

Portanto, o fato de não se considerar as *EDUs* encaixadas na análise implica uma perda considerável de granularidade na estrutura retórica do texto (Carlson and Marcu, 2001).

A estratégia de análise retórica usada na análise do corpus considera, assim, as *EDUs* encaixadas. A subseção a seguir comenta brevemente essa.

4.2 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 estruturas 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. A próxima seção apresenta um exemplo de análise discursiva.

4.3 Exemplo de Análise Discursiva

Considere o texto-exemplo 1 na Figura 5. O primeiro passo da análise é a segmentação do texto em *EDUs*. A primeira sentença do texto corresponde a uma única *EDU*, *EDU* [1]. A segunda sentença também corresponde a uma única *EDU*, *EDU* [2]. Já a terceira sentença apresenta duas *EDUs*, [3] e [4], sendo a *EDU* [4] sinalizada pelo marcador temporal "desde". A quarta sentença também apresenta duas *EDUs*, [5] e [6], sendo a *EDU [6]* uma oração complementar do verbo de atribuição "disse". A quinta sentença apresenta três *EDUs*, [7], [8] e [9], cuja *EDU* [8] é um complemento do verbo de atribuição "disseram" e a *EDU* [9] é sinalizada pelo marcador discursivo "para". Assim como na quarta sentença, na sexta sentença apresenta-se duas *EDUs*, [10] e [11], em função do verbo de atribuição "disse". Já a sétima sentença corresponde a uma única *EDU*, *EDU* [12]. Por fim, a oitava sentença corresponde a três *EDUs*, [13], [14] e [15], sendo a *EDU* [14] sinalizada pelo verbo de atribuição "disse" e a *EDU* [15] sinalizada por um relacionamento condicional com a *EDU* [14]. A Figura 6 apresenta o texto segmentado.

Aviões da Otan bombardearam ontem posições sérvias ao norte de Sarajevo. O ataque foi uma represália à tomada por sérvios de armamentos pesados, retirados da zona de exclusão da ONU em torno da cidade.

Foi a primeira ação da Otan contra sérvios desde o ataque aéreo às suas posições no encrave de Gorazde, em abril.

Um porta-voz do Departamento de Estado dos EUA disse que dois aviões norte-americanos e dois franceses bombardearam às 18h35 (13h35 em Brasília) posições sérvias ao redor de Sarajevo.

Porta-vozes militares disseram que um total de 12 aeronaves, com a participação também de holandeses, saíram de bases da Otan na Itália para realizar os bombardeios.

A ONU disse que após o ataque os sérvios se comprometeram a devolver imediatamente as armas.

O comandante do Exército sérvio garantiu ao comandante das tropas da ONU em Sarajevo, general Michael Rose, que todas as armas retiradas da zona de exclusão seriam devolvidas até hoje. Rose disse que se a promessa não for cumprida haverá novos ataques.

Figura 5: Texto-exemplo 1[7]

[1] Aviões da Otan bombardearam ontem posições sérvias ao norte de Sarajevo. [2] O ataque foi uma represália à tomada por sérvios de armamentos pesados, retirados da zona de exclusão da ONU em torno da cidade.

[3] Foi a primeira ação da Otan contra sérvios [4] desde o ataque aéreo às suas posições no encrave de Gorazde, em abril.

[5] Um porta-voz do Departamento de Estado dos EUA disse que [6] dois aviões norte-americanos e dois franceses bombardearam às 18h35 (13h35 em Brasília) posições sérvias ao redor de Sarajevo.

[7] Porta-vozes militares disseram que [8] um total de 12 aeronaves, com a participação também de holandeses, saíram de bases da Otan na Itália [9] para realizar os bombardeios.

[10] A ONU disse que [11] após o ataque os sérvios se comprometeram a devolver imediatamente as armas.

[12] O comandante do Exército sérvio garantiu ao comandante das tropas da ONU em Sarajevo, general Michael Rose, que todas as armas retiradas da zona de exclusão seriam devolvidas até hoje. [13] Rose disse que [14] se a promessa não for cumprida [15] haverá novos ataques.

Figura 6: Segmentação do texto-exemplo 1 em *EDUs*

Depois da fase de segmentação do discurso, o próximo passo no processo de análise é a identificação das relações retóricas entre pares de *EDUs* adjacentes em uma

[7] Extraído de um texto do corpus TeMário (Pardo e Rino, 2003).

mesma sentença (se existirem) e a especificação das *EDUs* nucleares e satélites, para a construção das subestruturas RST.

Considere as *EDUs* [3] e [4] na terceira sentença. O marcador "desde" em [4] sinaliza uma relação retórica *TEMPORAL-AFTER* entre as duas *EDUs*, pois a *EDU* [3], considerada a mais importante e, portanto, o núcleo, apresenta uma situação que ocorreu depois da situação apresentada em [4], o satélite. Na quarta sentença, o verbo de atribuição "disse" sinaliza uma relação *ATTRIBUTION* entre as *EDUs* [5] e [6], sendo a *EDU* [5] o satélite e a *EDU* [6] o núcleo. Na quinta sentença, o verbo de atribuição "disseram", na *EDU* [7], sinaliza uma relação *ATTRIBUTION* entre ela e as *EDUs* [8] e [9], que, por vez, relacionam-se através da relação *PURPOSE*, sinalizada pelo marcador discursivo "para", pois pode-se dizer que "as 12 aeronaves saíram de bases da Otan com o propósito de realizar os bombardeios". As *EDUs* [7] e [9] são os satélites e a *EDU* [8] é o núcleo. Na sexta sentença, as *EDUs* [10] e [11] relacionam-se por meio de uma relação *ATTRIBUTION*, também sinalizada pelo verbo de atribuição "disse", sendo a *EDU* [10] o satélite e a *EDU* [11] o núcleo. Na oitava sentença, o verbo de atribuição "disse", na *EDU* [13], sinaliza uma relação *ATTRIBUTION* entre ela e as *EDUs* [14] e [15], que, por vez, relacionam-se através da relação *CONDITION*, pois a situação apresentada em [14] é a condição para a realização da situação apresentada em [15]. As *EDUs* [13] e [14] são satélites e a *EDU* [15] é o núcleo. A Figura 7 ilustra as subestruturas RST de cada uma dessas sentenças. A subestrutura (a) corresponde a terceira sentença, a subestrutura (b) corresponde a quarta sentença, a subestrutura (c) corresponde a quinta sentença, a subestrutura (d) corresponde a sexta sentença e a (e) corresponde a oitava sentença.

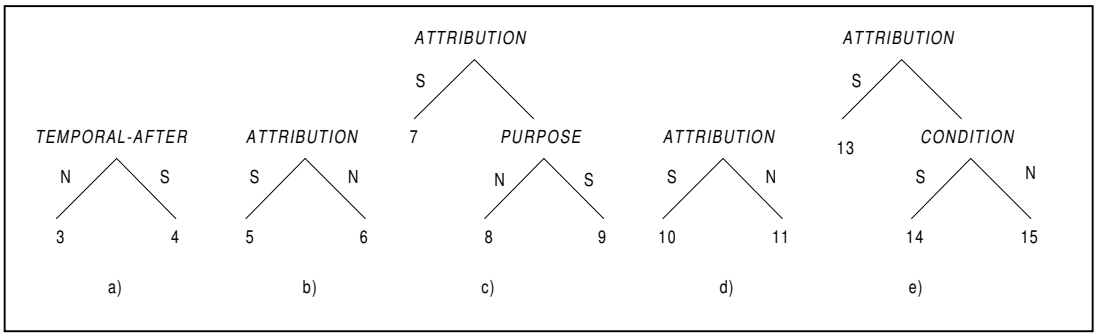

Figura 7: Subestruturas RST de cada sentença do texto-exemplo 1.

Uma vez construídas as subestruturas RST de cada sentença, relacionam-se as subestruturas de todas as sentenças de um parágrafo (caso haja mais de uma sentença). As duas sentenças no primeiro parágrafo do texto, as quais correspondem as *EDUs* [1] e [2], respectivamente, relacionam-se por meio de uma relação *REASON*, pois a *EDU* [2] apresenta a razão para o fato ocorrido na *EDU* [1], sendo [1] o núcleo da relação e [2] o satélite. No último parágrafo do texto, a primeira sentença correspondente a *EDU* [12] será relacionada à subestrutura da segunda sentença desse parágrafo, ou seja, à subestrutura (e) ilustrada na Figura 7. A relação se estabelece por meio da relação *ELABORATION*, pois a informação apresentada na segunda sentença elabora a informação apresentada na primeira sentença do parágrafo, sendo a primeira sentença o núcleo e a segunda o satélite. A Figura 8 ilustra as subestruturas de cada parágrafo do texto.

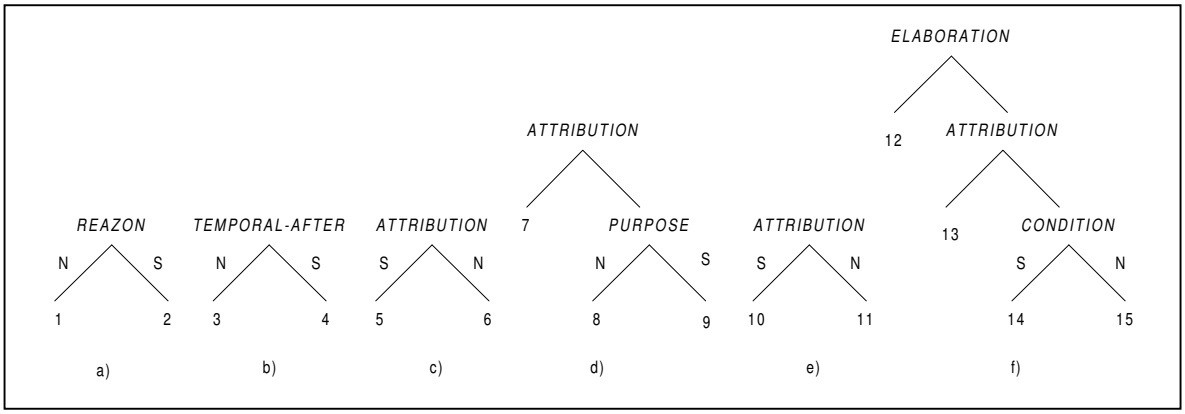

Figura 8: Subestruturas RST de cada parágrafo do texto-exemplo 1

Por fim, relacionam-se as subestruturas de todos os parágrafos do texto. As subestruturas (a) e (b) da Figura 8, referentes ao primeiro e segundo parágrafo, respectivamente, relacionam-se através da relação *ELABORATION*, pois a informação apresentada no segundo parágrafo, ou seja, nas *EDUs* [3] e [4], elabora a informação apresentada no primeiro parágrafo, isto é, em [1] e [2]. Assim, o parágrafo 1 é o núcleo e o parágrafo 2 é o satélite da relação. A Figura 9 ilustra a subestrutura resultante desse relacionamento.

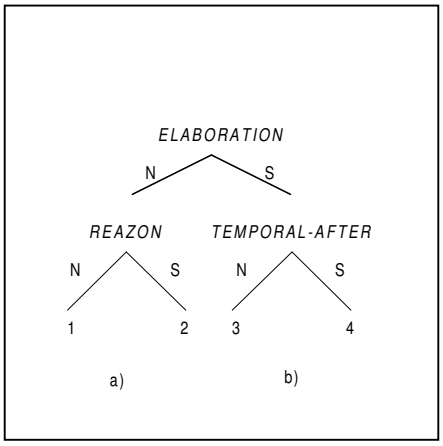

Figura 9: Subestrutura RST do primeiro e segundo parágrafo

Por vez, essa subestrutura relaciona-se com a subestrutura (c), correspondente ao parágrafo 3, por meio da relação *RESTATEMENT*, uma vez que a informação apresentada no parágrafo 3, ou seja, nas *EDUs* [5] e [6] reitera a informação apresentada no núcleo mais à esquerda dessa subestrutura, correspondente a *EDU* [1]. A subestrutura resultante desse relacionamento é ilustrada na Figura 10.

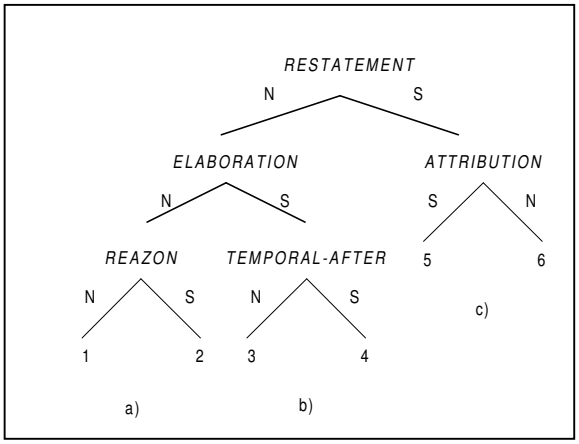

Figura 10: Subestrutura dos parágrafos primeiro, segundo e terceiro.

Dando continuidade à agregação dos parágrafos restantes, a subestrutura (d) elabora o fato apresentado pela *EDU* [1]. Assim, ela é agregada à subestrutura da Figura 10, produzindo a estrutura da Figura 11.

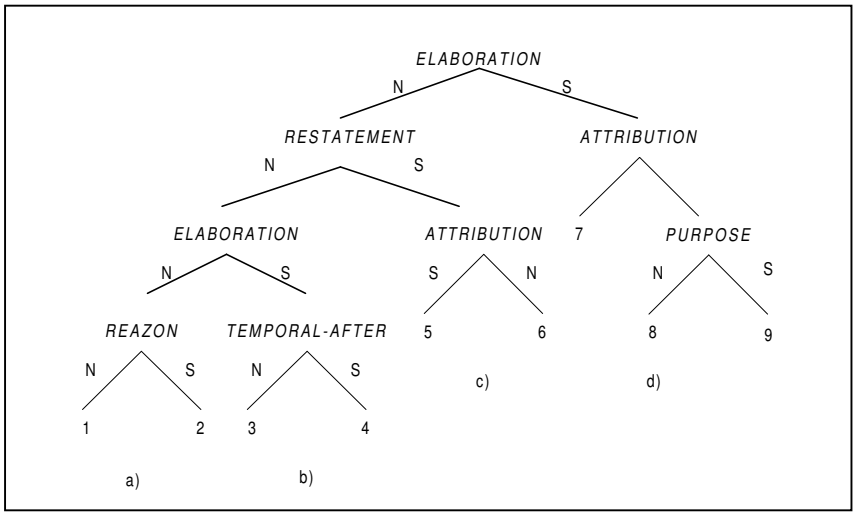

Figura 11: Subestrutura RST dos parágrafos primeiro, segundo, terceiro e quarto.

Analogamente, o quinto parágrafo elabora os parágrafos anteriores, como ilustra a Figura 12.

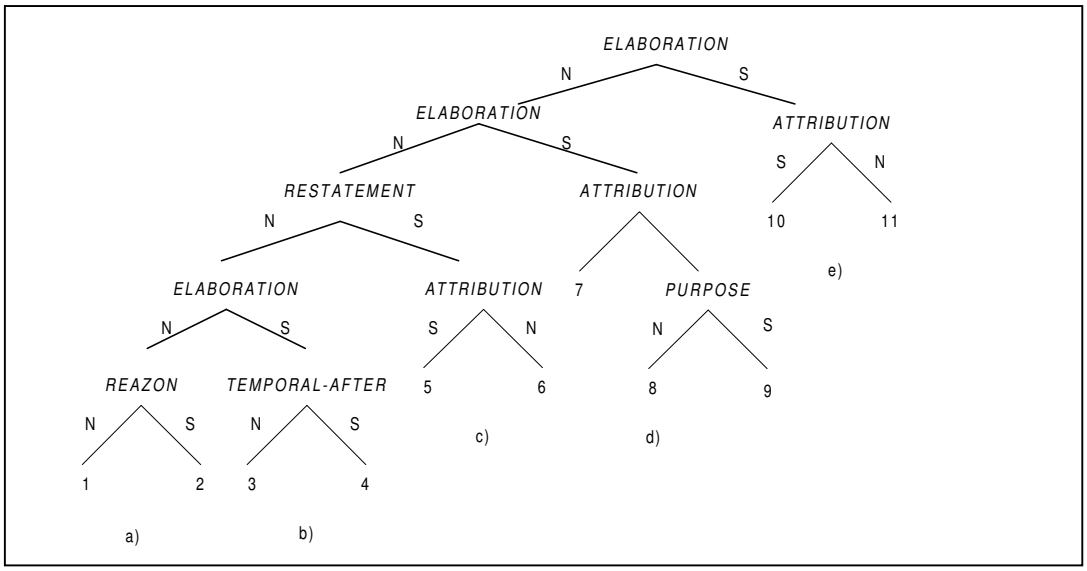

Figura 12: Subestrutura RST dos parágrafos primeiro, segundo, terceiro, quarto e quinto.

Por fim, como o sexto parágrafo (subestrutura (f), Figura 8) só se relaciona ao quinto parágrafo por meio da relação *EVIDENCE*, ele introduz uma nova subárvore como satélite da relação *ELABORATION* raiz da estrutura RST completa do texto, como mostra a Figura 13.

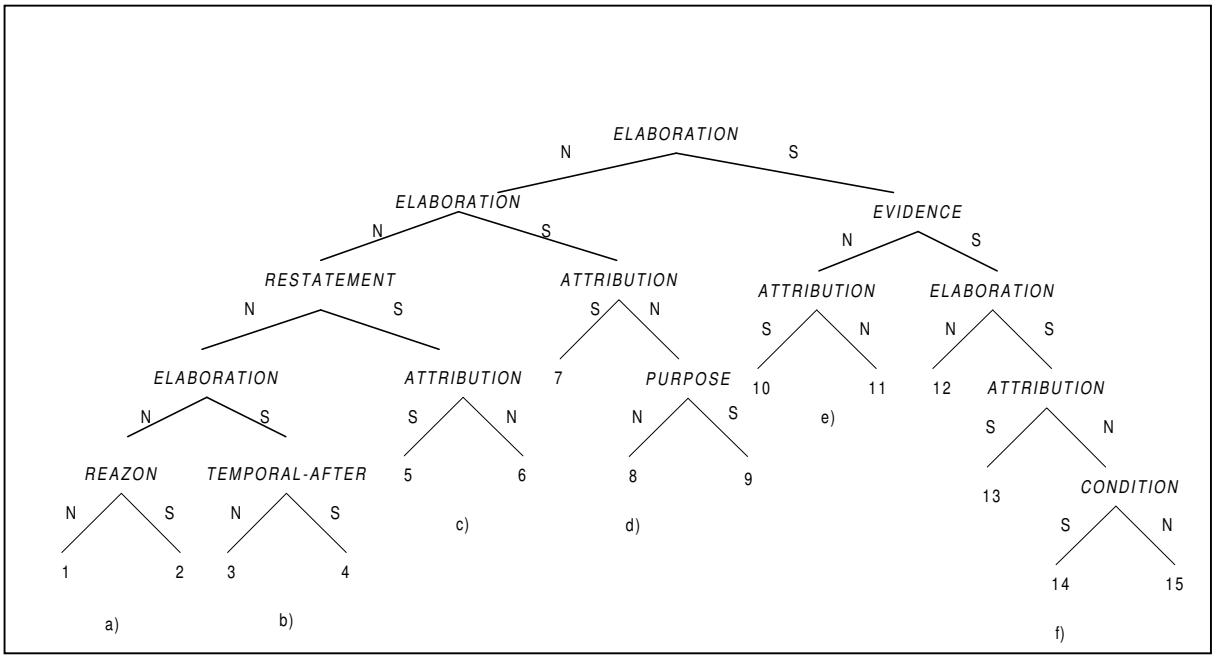

Figura 13: Estrutura RST do texto-exemplo 1

4.4 Conjunto de Relações Retóricas Considerado

Para a análise do corpus, inicialmente utilizou-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 and Marcu (2001). A Tabela 3 mostra o conjunto completo das relações utilizadas (as relações não definidas originalmente na RST são marcadas com "*").

Tabela 3: Conjunto de relações retóricas para análise do corpus

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

4.5 Síntese da Análise do Corpus TeMário

As tabelas 4 e 5 apresentam o número de ocorrências e a freqüência de cada relação retórica no corpus, sendo a Tabela 5 relativa somente às relações encaixadas. Como se pode notar na Tabela 4, algumas relações ocorreram com pouquíssima freqüência. Já a relação *ELABORATION* foi a mais freqüente no corpus. Isto talvez se justifique pela natureza do corpus: tratando-se de textos jornalísticos, elaborações sobre um mesmo tópico podem ser mais freqüentes do que em textos de outro gênero, por exemplo. A Figura 14 ilustra contextos de ocorrência dessa relação no segmento de discurso a seguir (extraído de um texto do TeMário):

"... [8] Descendente de africâners (colonizadores de origem holandesa), De Klerk nasceu em Johannesburgo, em 18 de março de 1936. [9] Seu pai foi membro do Partido Nacional (PN), fundado em 1948. [10] Foi o PN o principal responsável pela política do apartheid".

Tabela 4: Número de ocorrências e freqüência das relações retóricas

Relações Retóricas	**Ocorrência**	**Freqüência (%)**
ELABORATION	414	31.94
LIST	316	24.38
ATTRIBUTION	112	8.64
EVIDENCE	96	7.40
SAME-UNIT	89	6.86
SEQUENCE	76	5.86
PARENTHETICAL	13	1
REASON	36	2.77
JUSTIFY	32	2.46
CONTRAST	20	1.54
PURPOSE	17	1.31
JOINT	12	0.92
CONDITION	12	0.92
COMPARISON	9	0.69
CONCESSION	5	0.38
EXPLANATION-ARGUMENTATIVE	5	0.38
RESULT	5	0.38
CAUSE	6	0.46
TEMPORAL-AFTER	4	0.30
CIRCUMSTANCE	3	0.23
EXAMPLE	4	0.30
INTERPRETATION	3	0.23
MEANS	2	0.15
CONCLUSION	2	0.15
RESTATEMENT	1	0.07
BACKGROUND	1	0.07
TEMPORAL-SAME-TIME	1	0.07
Totais de ocorrências	1296	

Em relação às relações retóricas encaixadas (na Tabela 5), nota-se que a relação que ocorreu com maior freqüência foi a *ELABORATION-e*, devido à presença de muitas orações relativas no corpus. A Figura 15 ilustra contextos de ocorrência dessa relação no segmento de discurso a seguir (extraído de um texto do TeMário):

"[1] Os sul-coreanos tentaram ontem angariar o apoio mundial à sua denúncia contra a Coréia do Norte pela violação da trégua acertada em 1953, [2] que acabou com a guerra entre os dois países".

Tabela 5: Número de ocorrências e freqüência das relações retóricas encaixadas

Relações Retóricas	**Ocorrência**	**Freqüência (%)**
ELABORATION-e	140	77.34
CIRCUMSTANCE-e	8	4.41
EXPLANATION-ARGUMENTATIVE-e	6	3.31
PURPOSE-e	6	3.31
REASON-e	6	3.31
COMPARISON-e	4	2.20
JUSTIFY-e	3	1.65
MEANS-e	3	1.65
SUMMARY-e	1	0.55
ATTRIBUTION-e	1	0.55
CAUSE-e	1	0.55
CONCESSION-e	1	0.55
CONDITION-e	1	0.55
Totais de ocorrências	181	

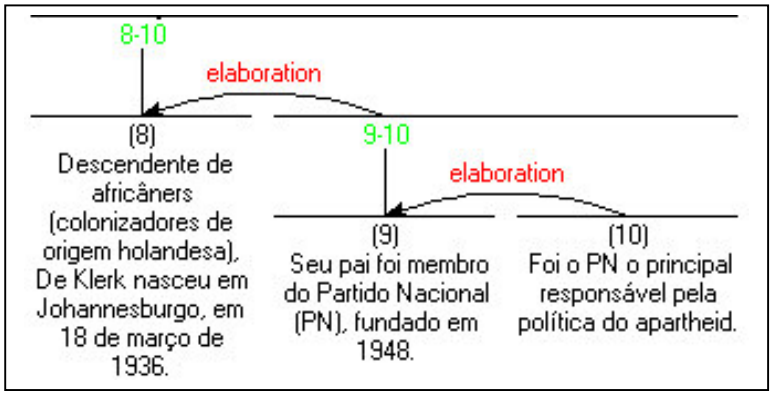

Figura 14: Exemplos de relação *ELABORATION*

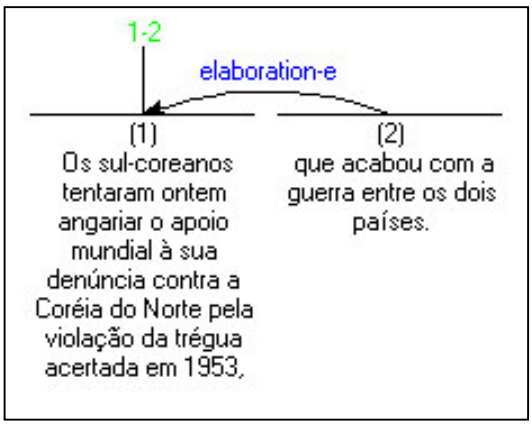

Figura 15: Exemplo de relação *ELABORATION-e*

4.6 Dificuldades da Análise Discursiva

A análise de discurso não é uma tarefa simples e fácil de ser realizada. Isto se deve, principalmente, à incompleta formalização da RST e a própria ambigüidade da língua. Por exemplo, na RST não há nenhuma formalização a respeito da segmentação do discurso, isto é, como identificar as *EDUs* em um texto. Além do mais, a RST contempla somente unidades discursivas e não unidades sintagmáticas e, contudo, utiliza-se das formas superficiais dos textos para determinar as *EDUs*.

A identificação das relações retóricas entre pares de *EDUs* adjacentes em um texto também não é uma tarefa fácil. A formalização de certas relações retóricas é obscura e algumas vezes leva a ambigüidade, isto é, pode haver casos onde mais de uma relação pode ser aplicada. Alguns marcadores discursivos também são ambíguos, por exemplo, o "*Porque*", que pode sinalizar uma relação *CAUSE, JUSTIFY, REASON*, entre outras. Por outro lado, a ausência de marcadores discursivos também dificulta a identificação da relação retórica. Nesses casos, o analista deve considerar o contexto e seu próprio conhecimento de mundo, para obter sucesso na identificação das relações.

O processo de determinação das *EDUs* nucleares e satélites também apresenta dificuldades, pois envolve a distinção, por parte do analista, entre o que o escritor do texto considera mais essencial para alcançar seu objetivo comunicativo e o que ele considera menos relevante. O sucesso dessa distinção depende, principalmente, da competência do analista como leitor. Em alguns casos a própria definição da relação retórica permite identificar as *EDUs* mais relevantes.

A RST também não provê nenhuma especificação formal sobre o critério de composicionalidade das subestruturas retóricas, ou seja, a agregação das subestruturas RST em uma única estrutura completa do texto. Esse critério de composicionalidade é crucial para a organização retórica textual (e não apenas proposicional), pois em um texto coerente há somente uma estrutura RST.

Em suma, o sucesso da análise de discurso depende não apenas do conhecimento do analista sobre a RST, mas, principalmente, da sua competência como leitor em utilizar sua própria experiência de mundo na interpretação e entendimento do discurso.

5. Considerações Finais

Este relatório apresentou uma análise discursiva segundo a *RST - Rhetorical Structure Theory* de um corpus de textos jornalísticos escritos em português do Brasil.

A análise do corpus consistiu no primeiro passo para a proposta de um modelo de produção de sumários a partir de estruturas RST de textos-fonte. O próximo passo a ser realizado consistirá na identificação das referências anafóricas e na identificação das características estruturais da organização RST em função dessas referências, para a reorganização estrutural RST, de sumários, a partir das estruturas RST de seus textos-fonte correspondentes.

Referências Bibliográficas

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

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. In I. Mani and M. Maybury editors, *Advances in Automatic Text Summarization,* pp. 15-21, The MIT Press.

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

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

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. *The Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization,* pp. 82-88. Madrid, Spain, July 11.

Marcu, D. (1999). Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury editors, *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.

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

O'Donnell, Michael (1997b). RST-Tool: An RST Analysis Tool. *Proceedings of the 6th European Workshop on Natural Language Generation*, March 24 – 26. Gerhard-Mercator University, Duisburg, Germany.

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

Pardo, T. A. S. (2002). *DMSumm: Um Gerador Automático de Sumários*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. São Carlos - SP.

Pardo, T.A.S. (2003). *DiZer: Discourse analyZER for Brazilian Portuguese*. Monografia de Qualificação ao Doutorado. ICMC/USP São Carlos-SP.

Pardo, T.A.S. e Nunes, M.G.V. (2003). *Segmentação Textual Automática: Uma Revisão Bibliográfica.* Série de Relatórios Técnicos: NILC-TR-03-02, ICMC/USP, São Carlos.

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.

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.