AuTema-Dis: uma arquitetura computacional para identificação da temática discursiva em textos em Língua Portuguesa

Ana Luísa Varani Leal

Orientador: Prof. Doutor Paulo Miguel Duarte Quaresma

Co-Orientador: Profa. Doutora Maria João Marçalo

Tese submetida à Universidade de Évora
para a obtenção do grau de Doutor em Informática

Departamento de Informática
Universidade de Évora

Setembro, 2008

Esta versão não apresenta os comentários do júri

AuTema-Dis: uma arquitetura computacional para identificação da temática discursiva em textos em Língua Portuguesa

Ana Luísa Varani Leal

Orientador: Prof. Doutor Paulo Miguel Duarte Quaresma

Co-Orientador: Profa. Doutora Maria João Marçalo

Departamento de Informática
Universidade de Évora

Setembro, 2008

Esta versão não apresenta os comentários do júri

Resumo

Os trabalhos desenvolvidos em Linguística Computacional direcionados à análise textual procuram explicar o comportamento e as relações de alguns fenômenos e mecanismos linguísticos, devidamente marcados na estrutura textual. Tais estudos revelam-se, na sua maioria, relacionados às questões de ordem morfológica, sintática e, conforme observamos, raras investigações avançam em direção à semântica. Em grande parte das investigações realizadas na área, o texto é o concreto objeto de estudo, a partir do qual os processos relacionados a sua constituição são explicados. A coerência textual emerge das relações retóricas que se produzem ao longo da configuração textual, ela é representativa da estrutura discursiva e, assim sendo, deve ser analisada e compreendida em termos globais. Para tal, é necessário considerarmos todos os níveis textuais relacionados e envolvidos no processo de significação. Neste sentido, desenvolvemos uma proposta metodológica para análise da temática discursiva. A proposta foi avaliada, nomeadamente, os processos localizados e a suas relações na constituição do tema, visando a produção automática de uma macroproposição/macroestrutura, tendo-se obtido resultados considerados satisfatórios.

Abstract

In Computational Linguistics, the works on textual analysis try mainly to explain behaviors and relations of linguistics components, properly identified in the textual structure. Such research is specially related to the morphological and syntactic structures. Few works are focused on the semantic analysis, where the text is the concrete object of study. From the textual analysis, the processes related to their constitution are exploited and justified. The textual coherence, which emerges from the rhetoric relations that occur in the textual configuration, is the representative element of the discourse structure. Therefore, it must be analyzed and understood in global terms. For this, it is necessary to consider all textual levels related on the process of signification. In order to exploit the textual coherence, we propose and develop a methodology for analysis of the thematic discursive. The proposal was evaluated to identify the localized processes and their relations in the constitution of the subject, aiming to produce an automatic macro-proposition and macro-structure. The results are promising, in terms of precision and recall.

Ao Universo: Divino e Humano

Ao meu marido Jean Phylipe pelo incondicional amor e apoio.
Aos meus pais amados, Helena e Diorico.
Ao meu orientador PQ e sua família.
Ao meu colega Luís Rodrigues – *o meu artista*.
A minha Especialíssima Amiga *Zezinha* – Maria José Gomes.
Ao meu amigo Pedro Salgueiro – *Anjo da Guarda*.
As minhas amigas Cássia T. Santos e Sirlei Mucke – *minhas F1*.
Aos meus amigos Portugueses e Brasileiros.
E ... ao meu Chiquinho Brazuca – *meu fiel escudeiro*.
À Universidade de Évora e ao Departamento de Informática por oferecerem as condições adequadas à realização deste trabalho.
À CAPES/MEC–Brasil - Coordenadoria de Aperfeiçoamento Pessoal de Nível Superior – pelo financiamento dos meus estudos.
À Universidade de Santiago de Compostela e ao Grupo de Pesquisa do Professor Pablo Gamallo – por oferecerem a oportunidade de testar o sistema AuTema-Dis nos corpora em Galego e Espanhol.

Índice

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Este trabalho apresenta uma nova proposta de arquitetura para análise automática de discurso. A pesquisa em questão direcionou-se, especificamente, à elaboração de um arquitetura computacional que, a partir da sua implementação em um sistema, efetuasse automaticamente análise de um texto em sua totalidade, evidenciando relações retóricas entre os seus constituintes e que, ao final do processamento, gerasse uma estrutura sintética representativa da macroestrutura temática do discurso. Assim sendo, a partir da arquitetura proposta construiu-se o protótipo AuTema-Dis, um analisador temático discursivo, que articula informações de cunho formais do tipo morfo-sintáticas e informações relacionais, isto é, relações semânticas.

1.1 Motivação

O estudo de um texto ou de um discurso requer do investigador um profundo olhar crítico no que se refere aos elementos, às estruturas e às relações que perpassam por toda a sua composição. Todavia, esse olhar crítico não se restringe unicamente à análise da estrutura formal, mas também às possibilidades de transformação que surgem pela maneira como essa estrutura textual é composta pelo produtor de acordo com as suas intenções e objetivos. Reconhecer a organização e o papel dos elementos linguísticos relativos à composição não é suficiente para compreender o tema apresentado em um determinado texto; é necessário reconhecer as relações significativas que se estabelecem entre esses elementos textuais, que estão a compor o sentido do discurso. Dessa forma, identificar a linha temática que percorre uma produção oral ou escrita, bem como, o sentido subjacente à estrutura discursiva são desafios para quem estuda texto ou discurso.

Conforme apresentado, a motivação para a realização deste trabalho surgiu a partir da observação de diferentes processos; um deles refere-se à necessidade da comunidade científica, que faz uso de textos em formato digital e recorre a bancos de dados através da Internet, em buscar mais rapidamente a informação tratada em um texto selecionado na web; outro ponto está relacionado à falta de clareza e precisão nos *abstracts* que apresentam, em alguns casos, estruturas dissociadas umas das outras, independente da relação significativa que as envolve; um outro ponto importante que motivou a realização desta proposta foi tornar mais ágil, dinâmica e precisa a elaboração de respostas em sistemas do tipo *pergunta/resposta.*

Assim sendo, a motivação para este estudo justifica-se na medida em que o resultado da investigação não se restringe apenas a utilização por pessoas do meio acadêmico, as quais investigam e manipulam dados textuais. O resultado do nosso estudo prevê uma aplicação em diferentes áreas de investigação que trabalham com a informação textual digitalizada em diferentes contextos, sejam contextos voltados à Linguística, à Informática ou em sistemas meramente informativos de domínio público, como por exemplo, pesquisas em bibliotecas, museus e sites de busca.

1.2 A Proposta

O projeto AuTema-Dis prevê uma intersecção de conhecimentos de áreas distintas, ou seja, Linguística e Informática. Trata-se da construção de uma arquitetura que, implementada computacionalmente, realiza a análise textual, considerando as informações mais relevantes dispostas na superfície de um texto, bem como, as relações de significação que se estabelecem entre os elementos linguísticos que a compõe. O objetivo do trabalho foi desenvolver uma base metodológica, cuja sua sistematização fosse capaz de:

- reconhecer a informação principal em um determinado discurso, considerando o resultado de uma análise sintática automática;
- reorganizar as estruturas relevantes ao tema em árvores de dependência dos segmentos – DTS [1];
- atribuir automaticamente algumas relações retóricas entre os segmentos e subsegmentos organizados nas DTS's;
- produzir automaticamente uma estrutura sintética em língua natural, a qual representa informação apresentada na estrutura discursiva, considerando os resultados das etapas

[1]DTS: Dependency Tree Segments, em Português, Árvore de Dependência dos Segmentos. O conceito foi desenvolvido nesta pesquisa para ser empregado no âmbito do AuTema-Dis.

anteriores do processamento.

O AuTema-Dis propõe a identificação automática da macroproposição [2] de um texto. Para a produção automática da macroproposição, o sistema utiliza, em uma das etapas do processamento, o resultado do analisador sintático *Palavras*[3] como fonte para a criação de regras, que auxiliam no processo de identificação e delimitação das fronteiras dos segmentos mais relevantes à composição do tema; a escolha deste analisador deve-se ao fato de se tratar de um dos melhores analisadores da Língua Portuguesa existente na atualidade. Essas regras são sistematizadas e fazem parte de uma das etapas a serem realizadas pelo AuTema-Dis, etapa que realiza a segmentação automática dos textos. As regras sintáticas de segmentação textual têm como objetivo orientar na seleção das proposições relevantes de um texto, bem como, a produção automática da macroestrutura representativa do tema do texto.

Em uma das etapas do processamento, a estrutura do texto é representada em um sistema arbóreo, identificado por *DTS*, o qual demonstra esquematicamente as características da automatização da análise sintática realizada pelo *Palavras*. Especificamente a esta investigação, a estrutura arbórea - **DTS** auxilia na identificação e sistematização de macroproposições localizadas ao longo da estrutura textual, além de organizar de forma hierárquica os segmentos e as relações que se estabelecem entre as estruturas do texto. No tocante à forma representativa das **DTS's**, verifica-se se é possível reconhecer o objetivo do autor através da análise das macroproposições localizadas, bem como, é possível, em alguns casos, identificar as relações retóricas entre os segmentos. Neste sentido, a representação arbórea simboliza concretamente as proposições do discurso.

Conforme mencionamos, as informações sintáticas obtidas a partir do resultado do *Palavras* são a base para as regras, no entanto, salientamos que o sistema AuTema-Dis recorre a outras características e mecanismos linguísticos presentes nos textos, tais como: marcadores discursivos, elementos relacionais conjuntivos, verbos e seus argumentos e, em especial atenção, às relações retóricas - *RST* [4]. Os elementos linguísticos são avaliados e classificados no sistema Autema-Dis considerando as informações provenientes do resultado da etapa apresentada pela *Palavras*.

Assim sendo, a investigação desenvolvida manipula informações de distintas áreas: a Informática e a Linguística, trata-se de um estudo interdisciplinar que discute pontos de inter-

[2]Macroproposição: a definição do termo é empregada aqui conforme [5]. Macroproposições representam o significado de parágrafos, seções de texto, grupo/conjunto de seções e eventualmente o papel do próprio texto. Elas são contrapartidas formais das noções intuitivas dos principais pontos e do resumo de um texto, representadas por macroestruturas.

[3]Analisador sintático automático para o Português, desenvolvido por Eckhard B. (2000).

[4]Relações Retóricas: conceito apresentado por Mann and Thompson 1987.

secção entre ambas. Desta forma, o escopo da nossa pesquisa não se restringe a uma ou a outra ciência, bem como, à aplicação dos resultados obtidos.

1.3 Arquitectura AuTema-Dis

No âmbito desta tese, desenvolvemos uma proposta de arquitetura para análise automática de discurso, a qual prevê, considerando-se a sua execução em sistema computacional, a realização das seguintes etapas: a análise automática pelo *Palavras*; a identificação automática dos segmentos; a representação arbórea - **DTS**; a identificação das relações retóricas entre os segmentos e a produção da macroestrutura/macroproposição temática. A organização esquemática de cada um dos componentes da arquitetura pode ser observada na figura 1.1.

1.4 Principais Contribuições do Trabalho

A tese desenvolvida contribui para a compreensão e entendimento de como se processa automaticamente a identificação estrutural/formal e a organização conceitual de textos escritos em Português Brasileiro e Português Europeu. Neste sentido, a relevância do trabalho se efetiva principalmente em relação à proposta metodológica empregada para realização do processamento automático de textos.

Em *lato sensu*, apresentamos abaixo as contribuições que se destacam na proposta em questão. A exposição das contribuições é feita sucintamente, sendo elas descritas em detalhe ao longo desta tese nos capítulos que a compõem.

- A Metodologia

  A metodologia utilizada no desenvolvimento desta pesquisa, explicitada no capítulo 4 desta tese, é constituída por etapas diferenciadas de análise textual, as quais representam hierarquicamente a estruturação formal e a estruturação conceitual de textos em Português Brasileiro *PB* e Português Europeu *PE*. No que se refere às etapas metodológicas, salientamos três atividades nucleares a serem implementadas e realizadas pelo sistema:

  1. segmentação automática dos textos em *PB* e *PE* em segmentos e subsegmentos, em conformidade com as características formais e conceituais dos *corpora*;

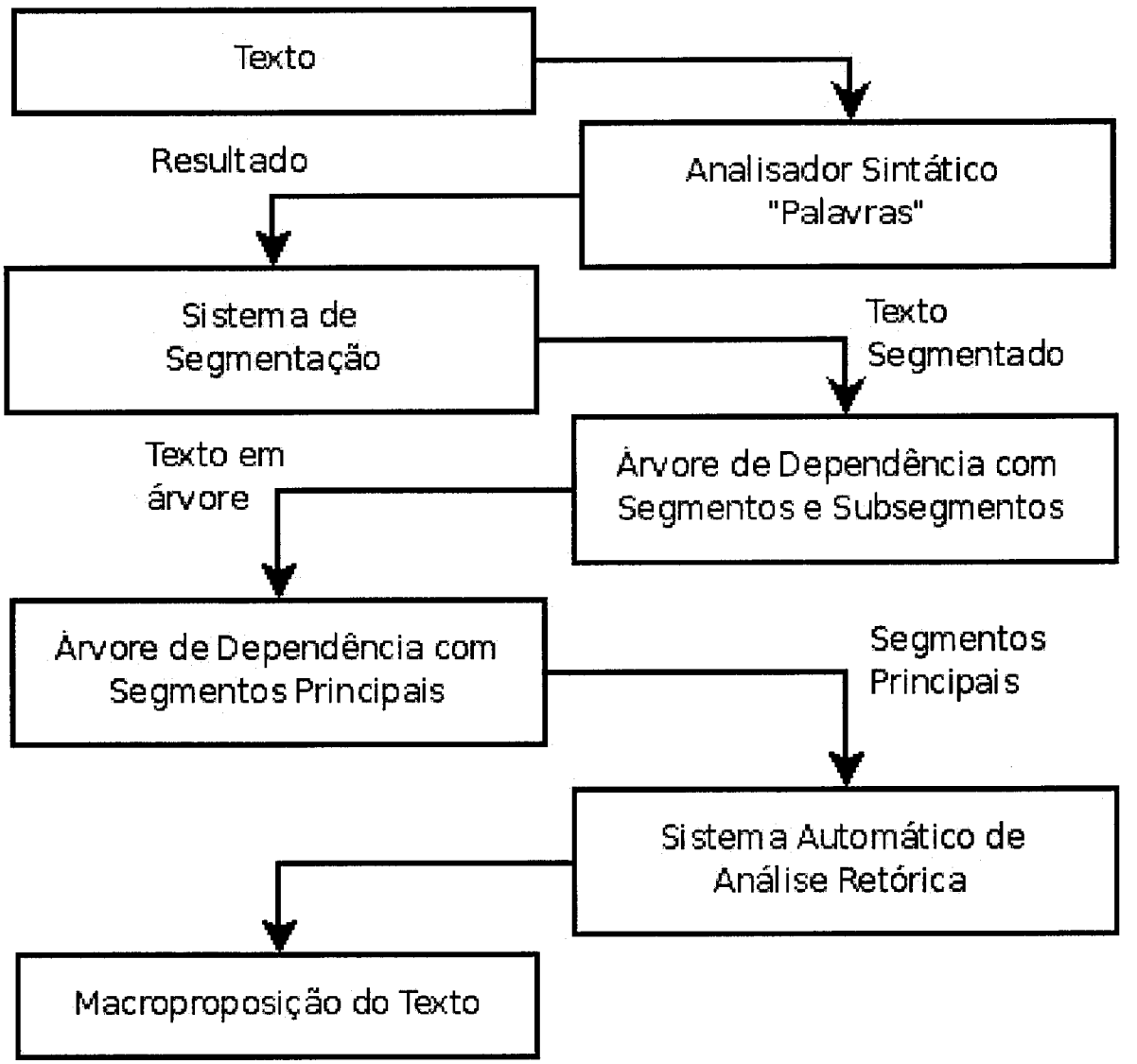

Figura 1.1: A figura representa de forma esquemática a arquitetura AuTema-Dis.

2. identificação e atribuição automática de algumas relações retóricas *RST* entre os segmentos e subsegmentos dos textos dos *corpora*;

3. identificação e produção automática de uma estrutura representativa da macro-proposição dos textos dos *corpora*, analisados pelo sistema.

Com base em testes realizados no sistema implementado a partir da arquitetura proposta, foi possível avaliar positivamente os resultados obtidos, conforme apresentamos no capítulo 5. Desta forma, estimamos ser a própria metodologia uma das contribuições significativas desta investigação. Assim, no que se refere à proposta em termos de análise discursiva, optou-se pela divisão do processo de investigação

em duas partes: uma parte que apresenta a nossa proposta teórico-metodológica em relação à identificação de uma estrutura representativa do tema do texto; e outra parte em que aplicamos a teoria desenvolvida, em um protótipo de sistema computacional, identificado pela sigla **AuTema-Dis** – Automatização da Temática Discursiva.

A investigação foi constituída por duas partes distintas: a primeira parte compreende a fundamentação teórica, levantamento dos dados, elaboração de regras, testagem e avaliação em *corpora* ; na segunda parte, a metodologia desenvolvida ao longo do percurso investigatório foi implementada e validada em sistema computacional.

- A Aplicabilidade do sistema teórico-computacional

  Como mencionamos, o trabalho desenvolvido é interdisciplinar e este fato justifica que os resultados e as conclusões contribuam em áreas distintas, isto é, Informática e Linguística. Em *lato sensu*, relacionam-se à aplicabilidade em áreas multidisciplinares, como por exemplo, relacionadas aos estudos puramente linguísticos e aos estudos computacionais de fenômenos linguísticos. Assim, salientamos que a aplicabilidade dos resultados obtidos na investigação pode ser constitutiva em pesquisas que envolvam:

  - sistemas computacionais capazes de recuperar informação contida em textos de forma automática;
  - sistemas inteligentes capazes de resolverem com situações de perguntas e respostas de maneira mais eficiente e rápida;
  - sistemas computacionais que realizam automaticamente a sumarização em textos em linguagem natural;
  - analisadores sintáticos automáticos – *parser* que realizam análise discursiva;
  - propostas teóricas Linguísticas com o auxílio de processamento computacional.

- Avaliação das Diversas Etapas Computacionais

  Conforme mencionado, a proposta metodológica para a realização desta tese foi desenvolvida em partes distintas, ou seja, uma parte teórica e outra parte computacional. A primeira parte foi a constitutiva teórica da proposta; a segunda parte, ou seja, a representativa computacional foi desenvolvida e testada para validar o constructo teórico proposto na primeira parte do trabalho.

  A proposta teórica é validada a partir da construção, implementação e execução do protótipo do sistema computacional AuTema-Dis. Todo o processo de avaliação e validação da componente teórica, bem como, da componente computacional é considerado como uma das contribuições deste trabalho, visto que cada uma das etapas

computacionais demonstra, na sua execução e no seu resultado, estar relacionada à teoria proposta.

O processo avaliativo das etapas constituintes do AuTema-Dis validam:

- a etapa do sistema que realiza a identificação e classificação dos segmentos e subsegmentos dos textos dos *corpora*;
- a etapa do sistema que executa a segmentação automática dos segmentos e subsegmentos dos textos dos *corpora*;
- a etapa do sistema de organiza em estruturas arbóreas os segmentos e subsegmentos dos textos dos *corpora*;
- a etapa do sistema que atribui automaticamente algumas das relações retóricas entre segmentos e subsegmentos dos textos dos *corpora*;
- a etapa do sistema que apresenta automaticamente a macroproposição de cada um dos textos dos *corpora*.

A avaliação foi realizada constante e sistematicamente através de estatísticas e testagem em *corpora*. Os resultados obtidos a partir da execução do sistema foram corroborados com os resultados da análise humana, o que confere ao processo avaliativo uma maior precisão e segurança nos resultados apresentados. Os dados estatísticos e os resultados são apresentados detalhadamente em tabelas no capítulo 5 desta tese.

1.5 Estrutura da Tese

O presente trabalho apresenta uma proposta metodológica para a análise automática de discurso e a sua implementação computacional. Em uma visão geral, resumimos a proposta à investigação ao texto jornalístico, para reduzir o escopo de análise. Embora a metodologia seja abrangente, acreditamos que para análise de textos mais específicos, ela deva ser ampliada e adequada.

Nesta introdução, discutimos a motivação para a realização deste estudo e a proposta metodológica que emerge do estudo em análise automática do discurso. Assim, apresentamos a arquitetura e a sua aplicabilidade teórico-computacional, bem como, os resultados obtidos com a sua implementação e execução em sistema automático. Além disso, discutimos as principais contribuições deste trabalho, as suas limitações e as possibilidades de ampliação.

O corpo desta tese está estruturado em torno da criação metodológica para análise automática de textos e da sua representação através da execução em sistema computacional. Deste modo, optamos pela divisão em duas partes: parte A, apresenta o processo teórico-metodológico, que serve de base para análise automática do discurso; enquanto que, na parte B, descrevemos a estruturação do sistema automático e apresentamos sua avaliação.

Cada uma das partes desta tese está estruturada de forma a demonstrar a metodologia e a sua aplicabilidade computacional. A parte **A** é composta pelos capítulos 2 e 3 e discute:

- Apresentação de trabalhos e teorias relacionados com análise automática do discurso;
- Discussão sobre teorias que sustentam as investigações que envolvem *texto* e *discurso*, em especial a *RST*; os mecanismos linguísticos que compõem a estrutura discursiva;
- Apresentação do enquadramento linguístico da proposta metodológica, elaborada para análise automática de discurso e arquitetura de para execução computacional;
- Descrição e caracterização dos *corpora*.

A parte B é constituída pelos capítulos 4 e 5 e discute:

- Apresentação da metodologia e representação da arquitetura;
- Descrição das estratégias de análise automática considerando as etapas do sistema computacional *AuTema-Dis*;
- Resultados e avaliações da proposta metodológica a partir da execução do sistema;

No capítulo 6, apresentamos as conclusões obtidas com a pesquisa, as aplicações e vantagens da metodologia enquanto sistema computacional; os avanços e as limitações, as possibilidades de ampliação em trabalhos relacionados.

Capítulo 2

Estado da Arte

Na sequência deste capítulo, apresentamos as teorias e trabalhos que apoiam em termos linguístico-computacionais a proposta em questão. Os estudos apresentados revelam-se de maneira significativa e profunda no tocante às pesquisas e investigações na área de análise automática do discurso, reconhecidos pela comunidade acadêmica.

Conforme mencionamos, no capítulo introdutório desta tese, o objetivo principal deste trabalho foi elaborar uma arquitetura que, a partir da sua implementação, realizasse automaticamente uma completa análise textual e que apresentasse, ao final do processamento, uma estrutura em língua natural representativa do tema presente no texto analisado. Face a tal objetivo, investigou-se uma teoria que fosse capaz de explicar não somente as relações formais de superfície textual, mas sim que explicasse e demonstrasse teoricamente as relações significativas entre os segmentos, que perfazem toda estrutura discursiva. Assim, optou-se por utilizar a *RST* - Teoria da Estrutura Retórica, apesar das suas limitações.

A *RST* foi originalmente desenvolvida pela equipe de pesquisadores da University of Southern California, William C. Mann (Information Sciences Institute) e Sandra A. Thompson (Linguistics Department) em 1983. O objetivo inicial da teoria era oferecer um suporte para os estudos em geração automática de texto, Geração de Língua Natural – *GLN*.

Desde a data inicial de publicação da RST até os dias de hoje, a teoria vem sendo ajustada, reestruturada e adaptada em muitos pontos, a fim de ser aplicada em diferentes estudos. Neste sentido, podemos identificar autores que têm trabalhado seriamente para ampliar e aprofundar os princípios estabelecidos pela RST, são eles Marcu Daniel [25; 26], Michael O'Donnell [34; 37], Thiago Pardo [38]. Cada um dos pesquisadores adequou os princípios da RST em conformidade à linha de investigação executada e tipo de abordagem pretendida com a pesquisa.

A utilização da *RST* tem sido valiosa como teoria, visto que, apresenta recursos para identificação da estrutura do texto, além da categorização de alguns fenômenos discursivos, podendo, em alguns casos específicos, oferecer uma explicação formal para os mecanismos e recursos intrínsecos à estrutura textual. O emprego da *RST* como ferramenta tem sido importante para o desenvolvimento dos estudos em *PLN*, principalmente aqueles que: analisam fenômenos como a coerência e a coesão; que executam processos de segmentação de textos, contribuindo nos processos de sumarização automática *SA*; que executam processamento e geração automática de textos em língua natural *GLN*; em estudos que relacionam o significado das conjunções e dos marcadores discursivos na composição da estrutura e sentido de um texto.

Independentemente da ênfase pretendida ou da adaptação que a RST tenha sido alvo, é fato que a teoria adquiriu um *status* importante em diferentes linhas de pesquisa, sejam elas relacionadas à Linguística, à Informática ou à Linguística Computacional. No caso da proposta em questão, a *RST* está sendo utilizada para auxiliar na identificação da cadeia temática através das relações entre os segmentos que compõem um texto e na edificação da macroestrutura/macroproposição em dos textos dos *corpora*.

2.1 Apresentação da *RST* – Teoria da Estrutura Retórica

Originalmente construída para suprir a carência teórica em relação aos estudos da geração de texto (GLN), RST - *Teoria da Estrutura Retórica – A Teoria da Organização do Texto*, foi desenvolvida inicialmente, em 1983, a fim de suprir tal lacuna. Conforme os autores, não havia até aquela data uma teoria que fosse capaz de explicar a função ou a estrutura discursiva de um texto e que pudesse fornecer uma orientação detalhada para a programação de um sistema de geração automática de um texto.

Da busca em descrever e explicar os processos que organizam a informação discursiva, os autores, William C. Mann e Sandra Thompson, propuseram a *RST* como uma teoria que explica texto [1], assim sendo, identificam-no como uma unidade sem ausência de lacunas e de conjuntos aleatórios de frases. Neste sentido, eles explicam a coerência dos textos - mais do que os processos que subentendem sua criação e sua interpretação. Para os pesquisadores, para toda parte de um texto coerente existe uma função, uma razão plausível

[1] Texto: em nossa investigação, texto é identificado como a realização superficial de discurso, neste sentido, reconhecemos discurso como uma complexa entidade que contém proposições em que o autor expressa os seus objetivos. O texto é o discurso realizado concretamente.

para sua presença, seja ela evidentemente clara para os leitores, isto é, que o leitor não tenha a sensação que faltem partes ao conjunto.

Trata-se de uma teoria descritiva que, a partir da organização dos discursos, demonstra as relações que se estabelecem entre suas estruturas em termos funcionais, identificando o ponto da relação e a extensão dos itens relacionados. A fim de melhor explicar os procedimentos textuais e a organização discursiva, os autores apresentaram inicialmente 25 relações retóricas na primeira versão da teoria, no entanto, estas relações, com a evolução das investigações foram ampliadas para o número de 32, as quais encontram-se nos anexos desta tese.

Conforme Mann e Thompson [24], avaliando-se essas relações pode-se evidenciar a base funcional da hierarquia textual, o que torna possível explicar a organização estrutural, bem como, a coerência. A teoria *RST* postula um conjunto de possibilidades que identificam e categorizam padrões estruturais, os quais se edificam entre as proposições que compõem um discurso, é neste sentido que a proposta teórica dos autores explica a organização textual. As relações retóricas cumprem as formalidades textuais básicas, para que o objetivo proposto pelo autor seja reconhecido pelo leitor.

A idéia central para a RST é a noção de *Relação Retórica* que se estabelece entre duas proposições discursivas (spans). Tais proposições (spans) desempenham diferentes *papéis ou funções*, ou seja, um é N - núcleo e o outro S - satélite. O núcleo representa o segmento mais significativo na relação, enquanto que o Satélite representa um conteúdo adicional em relação a esse Núcleo. Normalmente, as relações retóricas se edificam entre os pares de proposições (spans) com um Núcleo e um Satélite, trata-se de uma relação simples, denominada *Nuclear*, no entanto, esse não é o único tipo de relação que pode ser observado. Podemos identificar uma relação do tipo *Multinuclear*, em que se edificam relações entre spans de mesmo valor (núcleos) o que lhes confere uma relação retórica específica, conforme pode ser observado na figura 2.3, na qual apresentamos a relação retórica sequência na continuidade deste capítulo.

A identificação dos pares de proposições (spans), bem como, a atribuição dos papéis desempenhados por cada um deles está condicionada ao *valor significativo* que ambos ocupam no segmento discursivo do qual fazem parte. A determinação dos papéis/funções está sujeita às especificidades da pessoa que executa a análise, ou seja, do analista.

As estruturas apresentadas na figura 2.1 mostram os segmentos N e S em que podemos observar uma relação composta por apenas um núcleo. Os arcos com as setas apontam sempre em direção ao núcleo e a linha vertical apresenta o ponto nuclear na proposição.

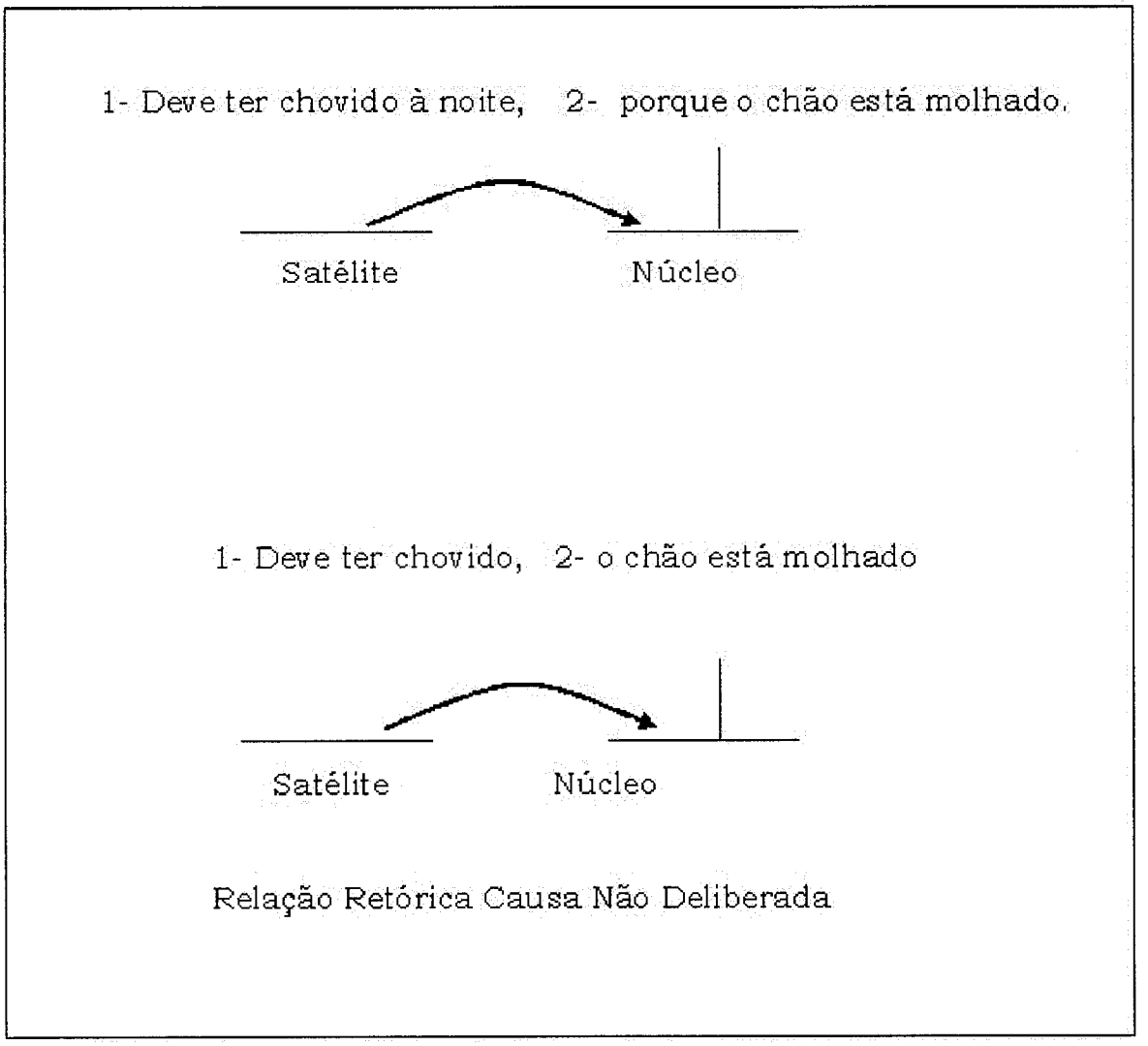

Figura 2.1: O exemplo apresenta a orientação de uma relação nuclear, relação retórica – causa involuntária.

Em relação aos elementos adotados pela RST, além das proposições (spans) Núcleo e Satélite, a teoria envolve outros elementos que participam na identificação das relações, são eles: W - *writer* - *escritor* - quem produz o texto; R - *reader* - *leitor* - quem recebe e interpreta o texto; e o A - *analyst* - *analista* - quem realiza a análise do texto. Dependendo do tipo de análise que se deseja realizar no texto, o analista elege um desses elementos, a fim de direcionar o tipo de análise.

Apesar da necessidade da participação efetiva de ao menos um desses elementos no processo de análise e na identificação das relações retóricas ao longo de todo texto, observa-se que tais atividades estão condicionadas ao olhar subjetivo que o analista manifesta sobre o texto. Ressalta-se este fato, pois é o analista a figura responsável pelos julgamentos a respeito

da composição; é ele quem condiciona a análise de acordo com as suas necessidades e prioridades em relação ao tipo de investigação que pretende realizar.

Todavia, a RST define quatro tipos de objetos de análise independentemente do tipo de texto ou linguagem:

- Relações;
- Esquemas;
- Aplicação de esquemas;
- Estruturas.

Conforme mencionamos, as relações retóricas ocorrem entre duas proposições (spans) - *ininterruptos intervalos lineares de texto* em uma unidade textual. Neste sentido, os esquemas são estabelecidos a partir das relações, definindo os arranjos constituintes da estrutura do texto. Em acréscimo a essa característica, os esquemas são responsáveis pela determinação de padrões pelos quais uma determinada proposição ou span de texto pode ser analisada em contraste com outras. Observa-se, desta forma, que são os esquemas que determinam como as proposições ou spans podem co-ocorrer. No que se refere à aplicação do esquema, são definidos padrões em que um esquema específico pode ser instanciado. No caso das estruturas, essas são compreendidas em relação ao texto completo, definidas em termos de aplicação da composição de um esquema, ou seja, é somente a partir da aplicação de um conjunto de esquemas que se pode representar uma estrutura discursiva completa.

Para os autores da *RST*, Mann e Thompson [24], as relações são definidas a partir da ligação que se estabelece entre duas proposições ou spans **N-S** ou **N-N** não sobrepostos. Todavia, para que se possa definir adequadamente uma relação entre duas proposições, faz-se necessário considerar algumas informações advindas de quatro áreas ou campos específicos, são eles:

- Restrições em relação ao núcleo.
- Restrições em relação ao satélite.
- Restrições na combinação de núcleo e satélite.
- Efeito.

Cada uma das quatro áreas ou campos representa uma determinada particularidade que deve ser observada entre duas proposições, a fim de que se possa determinar a manifestação de uma relação retórica entre elas. Essas quatro áreas ou campos estabelecem restrições relativas a cada uma das relações, são elas que determinam qual é a relação que pode ser estabelecida entre duas proposições, em que circunstâncias pode ser efetivada, a partir das suas características específicas. No caso do efeito, especificamente, o analista avalia se o que é apresentado ou desejado pelo escritor *w* é ou não plausível de ser realizado. Percebe-se, desta forma, que apesar das restrições preestabelecidas para cada uma das relações retóricas em questão, a análise e a classificação das relações estão subordinadas ao caráter subjetivo da análise realizada por um analista.

No entanto, apesar da subjetividade e das características peculiares a cada avaliação, as análises empiricamente realizadas por diferentes analistas, quando confrontadas, apresentam equivalência no que se refere à identificação e à classificação das proposições (spans), bem como, a determinação das relações retóricas estabelecidas entre elas, conforme foi observado e avaliado estatisticamente no âmbito desta investigação. O resultado da categorização das relações, justifica-se pela existência dessas condições preestabelecidas, que restringem à aplicabilidade de uma determinada relação. Assim sendo, para que uma relação possa ser atribuída entre duas proposições as condições preestabelecidas, que definem essas relações, devem ser cumpridas e identificadas para validar a classificação.

Neste sentido, como foi apresentado, para que cada uma das relações possa ser reconhecida como uma ocorrência válida entre duas proposições spans, devem ser consideradas as informações/restrições das áreas referidas acima, conforme o exemplo apresentado por Mann e Thompson [24] para a relação retórica Condição, mostrada na Tabela 2.1.

Os autores ponderam que a definição de um esquema seja proposta em termos de relações, sendo estas constituídas por padrões abstratos, dos quais fazem parte um pequeno número de proposições spans. Devido a essa caracterização, os esquemas são capazes de determinar as disposições dos constituintes na estrutura textual, podendo explicar como as proposições podem co-ocorrer. É nesse sentido que a aplicação de um esquema com determinadas condições é capaz de apresentar a estrutura retórica de um texto. Mann e Thompson [24] definem cinco tipos de esquemas, através dos quais as relações retóricas se organizam, conforme mostra a representação na figura 2.2.

O padrão estrutural mais observado é o Nuclear em que identificamos a presença de um núcleo e um satélite. No entanto, nem sempre é possível reconhecer tal situação, quando esse padrão não é identificado, podemos fazer uso de um tipo de esquema mais específico,

Elemento de definição	Ponto de Vista do Observador
Restrições no Núcleo N	Nenhuma
Restrições no Satélite S	S apresenta uma situação hipotética, futura ou não realizada (relativa ao contexto situacional).
Restrições na contribuição N + S	A realização da situação apresentada no N depende da realização do que foi apresentado em S.
O Efeito pretendido pelo leitor em usar a relação endereçada ao leitor L nunca é nulo.	O Leitor reconhece que a realização da situação apresentada no N depende da realização da situação apresentada no S.
Local do Efeito De onde o efeito é derivado.	N e S

Tabela 2.1: A tabela exemplifica uma das relações apresentadas por Mann e Thompson [24] com as restrições por área. Trata-se de um exemplo representativo da Relação Retórica de Condição.

isto é, o *multinuclear*. O esquema multinuclear é empregado sempre em que se observa uma possibilidade diferente na organização textual que não seja a nuclear (N - S). É graças ao esquema multinuclear que podemos justificar e avaliar padrões textuais em que estão presentes em outras formas de organização, como podemos observar na figura 2.3.

Para realizar as análises e determinar as relações existentes entre as estruturas, os autores observaram o tipo de organização e, a partir dessa organização, estabeleceram uma metodologia específica, que orienta para duas condições que devem ser observadas no momento de análise, são elas:

- Seccionar o texto em unidades de tamanho arbitrário, podendo ou não considerar alguma teoria para classificação;
- As unidades devem ter integridade funcional independente umas das outras, podendo ser orações ou cláusulas com exceção às orações subjetivas; às orações com função de complemento; à oração relativa restritiva, pois essas fazem parte do grupo das orações principais.

Para Mann e Thompson [24], uma completa análise estrutural de um texto é, por sua vez, a aplicação de um esquema, o qual restringe a sua aplicação em termos de organização da estrutura, obedecendo aos seguintes padrões:

- Completude (Completeness): a aplicação de um esquema contém o texto completo;

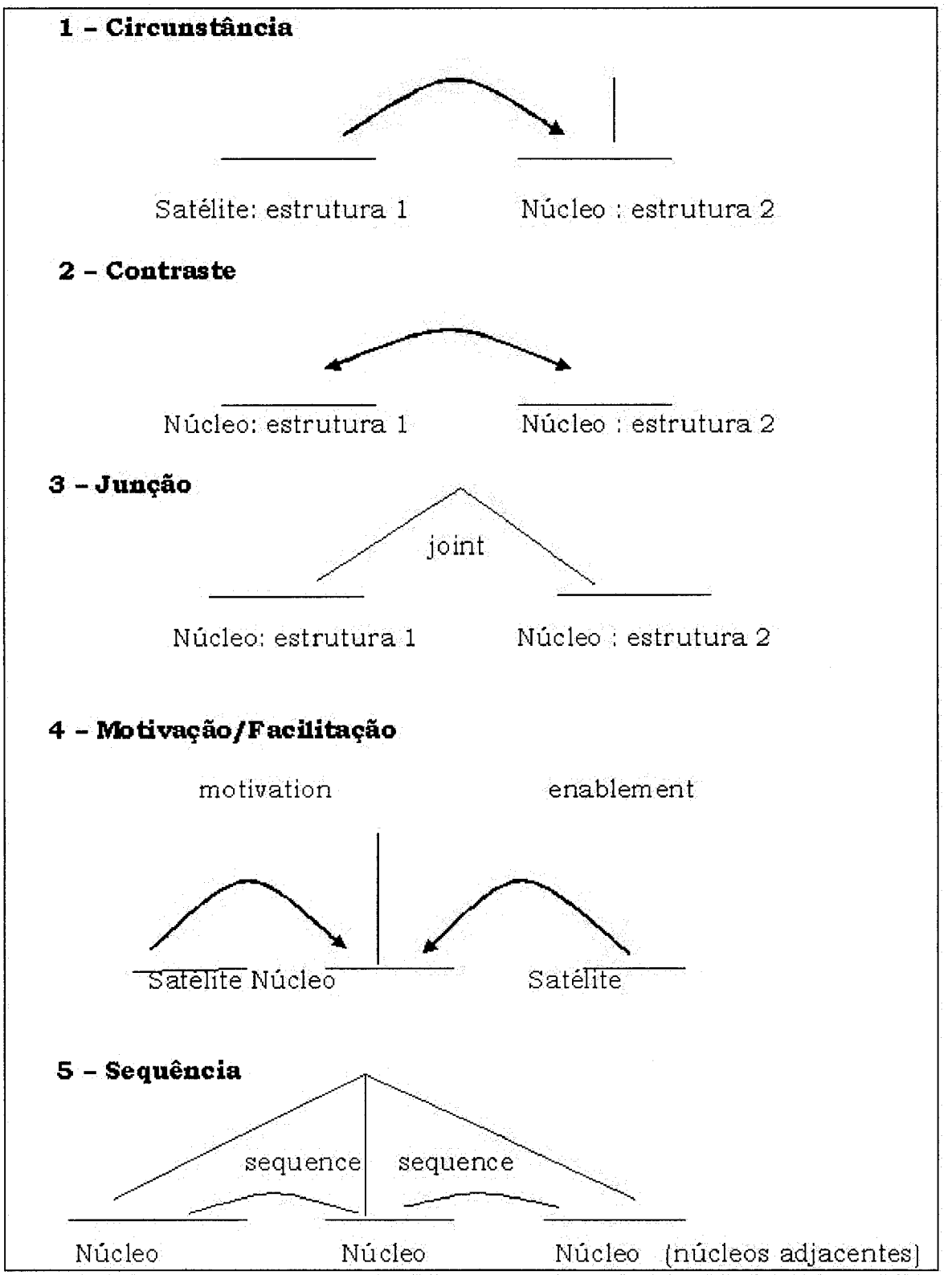

Figura 2.2: Apresentação dos 05 esquemas propostos para a organização das Relação Retóricas, conforme Mann e Thompson [24].

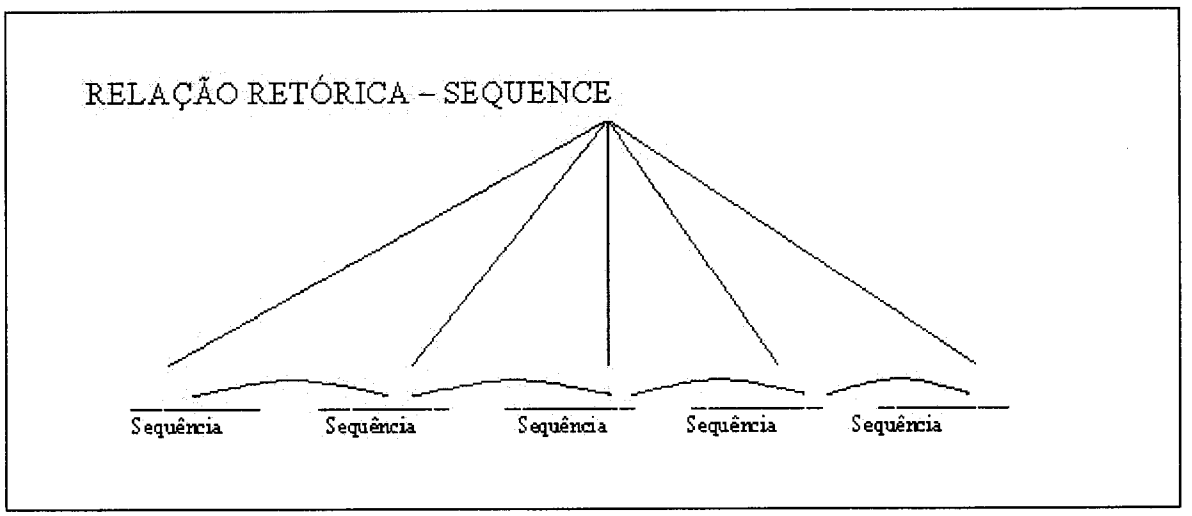

Figura 2.3: A relação retórica *sequence* ou *sequência* é um exemplo característico das relações retóricas multinucleares.

- Conectividade (Connectedness): cada proposição ou span, com exceção ao segmento ou span que contém o texto inteiro, é uma unidade mínima ou um constituinte da aplicação de outro esquema;
- Unicidade (Uniqueness): cada aplicação de um esquema envolve um conjunto diferente de segmentos/spans;
- Adjacência (Adjacency): as proposições/spans de cada aplicação de um esquema constituem uma contígua extensão de texto.

Conforme mencionado, o padrão estrutural das relações retóricas é composto por dois segmentos de texto religados de tal forma que um dos dois desempenha um papel específico em relação ao outro, ou seja, um é subordinante e o outro subordinado. Esta característica reforça a identificação das relações hipotáticas e paratáticas, reconhecendo-se neste contexto as relações nucleares e relações multinucleares. Outra característica importante a ser ressaltada está relacionada à organização da estrutura, isto é, a ordem dos segmentos não é precisamente determinada, mas para cada relação a ordem pode ser identificada como pares de proposições.

Reconhecendo a existência de relações entre os segmentos do texto e a fim de explicá-las, Mann e Thompson [24] definiram inicialmente pelo nome 24 relações retóricas e mais um esquema, sendo ampliadas mais tarde, conforme Taboada e Mann [41] para o número de 30 relações. As relações que se estruturam entre esses pares de proposições são classificadas de forma simples, sendo organizadas em dois grupos específicos de acordo com as suas semelhanças. Os dois grupos representam aspectos da estrutura do texto e é a partir da

distinção aspectual que esses dois grupos são propostos como: *subject matter* - relações que apresentam parte do conteúdo do texto e *presentational* - relações que são utilizadas para auxiliar na apresentação do processo.

A classificação em um ou em outro grupo está subordinada à interpretação do analista ou do leitor e ao efeito que uma determinada relação lhe causou. Abaixo, apresentamos a tabela 2.2 na qual estão contidas as relações definidas pelos autores com a inclusão das novas relações retóricas e esquema. Salientamos que as relações estão reunidas em conformidade com o tipo de relação que desempenham e se representam uma relação multinuclear.

Relações de Conteúdo	**Mult.**	**Relações de Apresentação**	**Mult.**
1- Circunstância	Não	19- Antítese	Não
2- Condição	Não	20- Segundo Plano/Fundo	Não
3- Elaboração	Não	21- Concessão	Não
4- Avaliação	Não	22- Capacitação	Não
5- Interpretação	Não	23- Evidência	Não
6-Causa involuntária	Não	24- Motivação	Não
7-Resultado involuntário	Não	25- Justificação	Não
8- Alternativa (Otherwise)	Não		
9- Propósito	Não		
10- Reformulação	Não		
11- Solução	Não		
12- Resumo	Não		
13- Causa voluntária	Não		
14- Resultado voluntário	Não		
15- Contraste	Sim		
16- Junção/União	Sim		
17- Sequência	Sim		
18- Lista	Sim		

Tabela 2.2: A tabela original apresenta as Relações Retóricas propostas por Mann e Thompson [24], conforme http://www.sfu.ca/rst/07portuguese/intro.html.

Na versão revisada da teoria em 2005, Taboada e Mann [41] ampliam para 30 as relações, mantendo a classificação original de acordo com a natureza de cada relação, ou seja, de conteúdo ou de apresentação, para as relações entre Núcleo e Satélite, denominadas de *Nucleares*. As relações núcleo-satélite de conteúdo tem como objetivo que o leitor reconheça a relação em causa. No caso das relações de apresentação, o objetivo concentra-se em aumentar a posição tendencial do leitor em relação ao que é proposto. O conjunto das relações pode ser evidenciado na figura 2.4 e na figura 2.5. Em referência às relações multi-nucleares, os autores preocuparam-se ampliando o rol de 4 para 7 relações e esquemas, conforme figura 2.6.

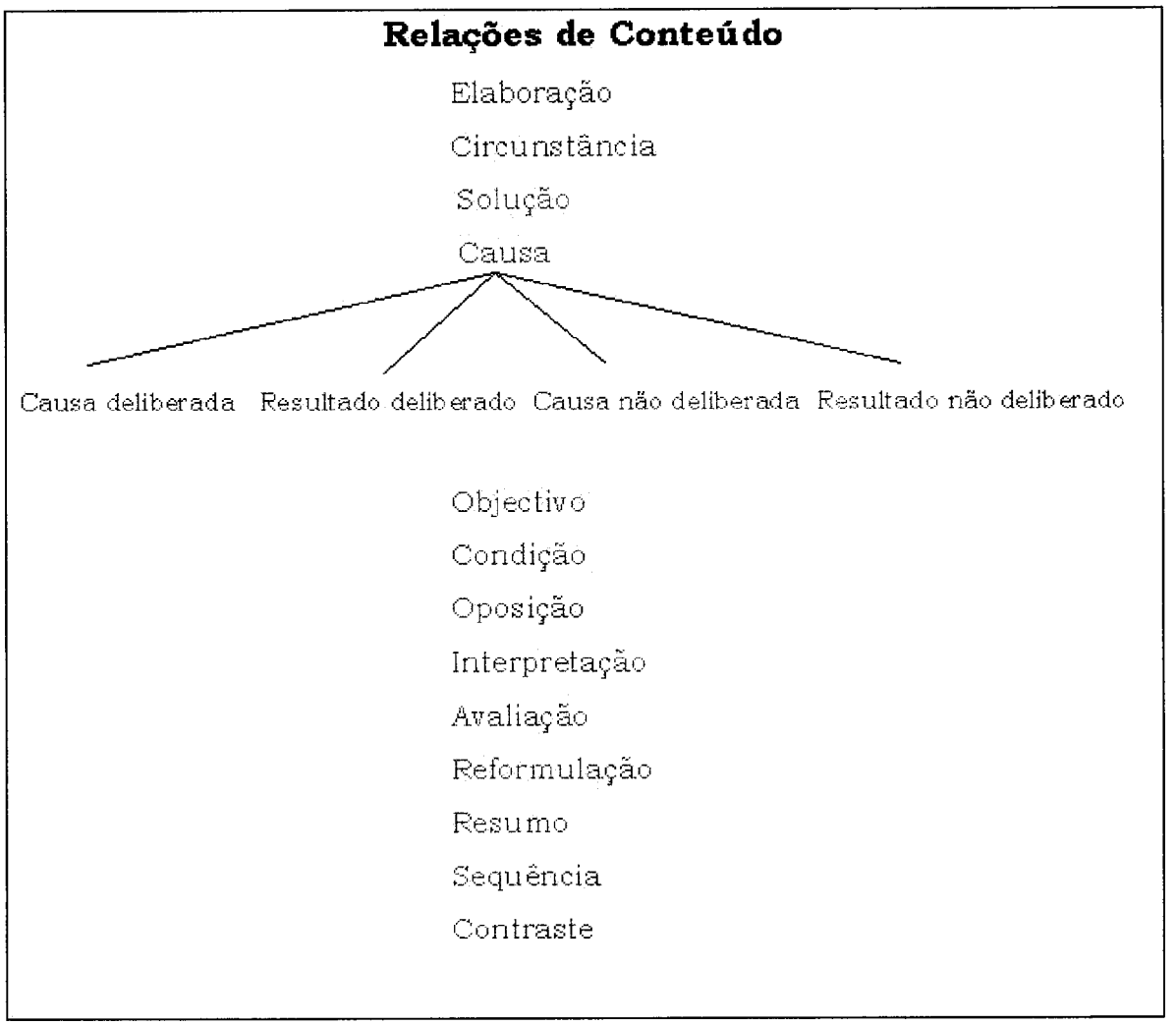

Figura 2.4: As relações retóricas de conteúdo podem ser ou não multinucleares, Mann e Thomposn [24].

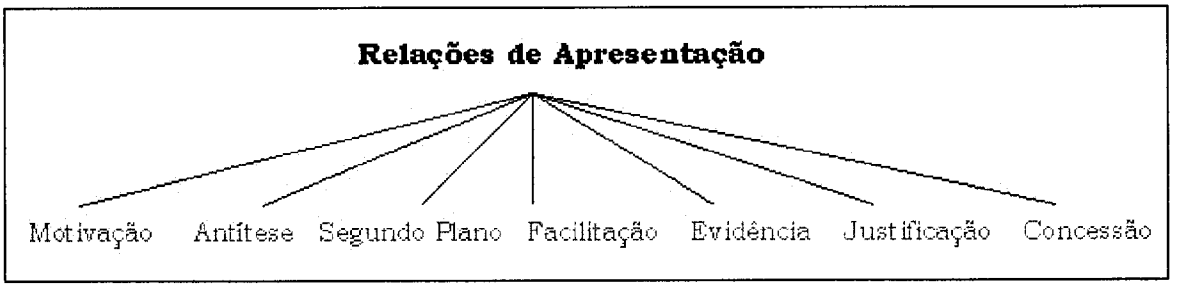

Figura 2.5: As Relações Retóricas de Apresentação Mann e Thompson [24].

Para cada uma das relações apresentadas na tabela tabela 2.2, os autores apresentam distinções que devem observadas no momento em que o analista realiza classificação das relações no corpus analisado. As características não estão relacionadas especificamente ao

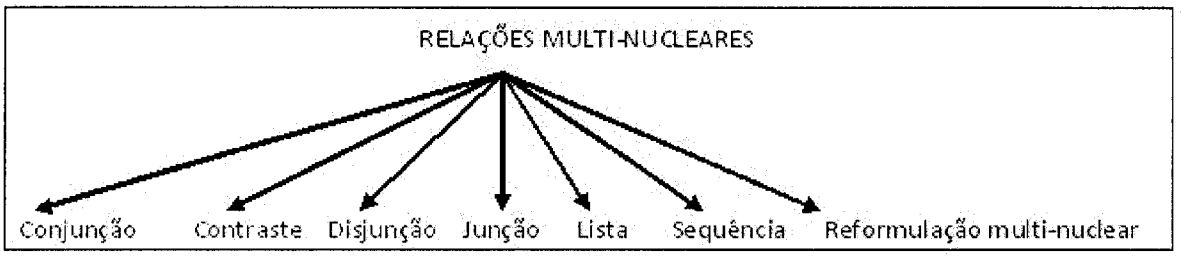

Figura 2.6: Relações retóricas multinucleares ampliadas por Mann e Thompson [24].

nome da relação, mas à restrição sobre Núcleo; à restrição sobre Satélite; à restrição do Núcleo e do Satélite em simultaneidade; ao Efeito e ao Local do Efeito, conforme apresentado na tabela 2.1. Essas restrições são propostas para auxiliar o analista a não fazer uso inapropriado ou inadequado de uma relação.

A fim de apresentar mais claramente as relações, os autores Mann e Thompson [24] elaboraram uma tabela em que podemos observar a definição para cada uma das relações retóricas, bem como, reconhecer as restrições que envolvem cada uma das proposições, sejam estas proposições *núcleo* ou *satélite*, as restrições que surgem da união *N e S* e como se manifesta a intenção do produtor do texto. A tabela em questão encontra-se nos anexos desta tese na versão em Português, obtida na própria página da teoria retórica.

A maneira como um analista elege as proposições de um texto , bem como, realiza a segmentação e a categorização das relações retóricas existentes entre suas proposições é delimitada em conformidade com o tipo de investigação que deseja realizar e com o tipo de aplicação que se pretende com os resultados obtidos com a análise. Essa flexibilidade na eleição das proposições, em alguns casos, pode gerar um certo tipo de ambigüidade nas análises, a qual pode ser evidenciada pela variabilidade nos resultados. Além disso, a seleção, a análise e a categorização dos proposições estão subordinadas à ação e ao julgamentos do analista, o que subentende-se por um processo subjetivo. No entanto, o trabalho de investigação realizado por Mann e Thompson [24] aponta para uma possibilidade que pode ajudar a delimitar de forma mais concreta e precisa as análises e a identificação das relações, afastando-se da eminente subjetividade.

Nesse sentido, os autores chamam a atenção para diferença na ordenação das proposições nas estruturas discursivas, demonstrando que existem relações tais como *antítese, fundo (segundo plano), concessão, condição, justificação e solução* que apresentam o Satélite posicionado antes do Núcleo; e as relações de *elaboração, facilitação, evidência, propósito (objetivo) e reformulação* apresentam o Núcleo antes do Satélite. A ordem na disposição das proposições nas estruturas, como observam os autores, parece ser independente ao controle

do escritor. No entanto, conforme os teóricos, é possível identificar um padrão na disposição das estruturas e relacionar este padrão a um tipo específico de relação, fato que auxilia na classificação das proposições.

Apesar de podermos contar com as informações sobre a ordem (posição) das proposições *N* e *S* em um texto, com a finalidade de restringir a ocorrência de uma específica relação, Mann e Thompson [24] chamam a atenção para a possibilidade de serem produzidas múltiplas análises, fato que vem a ser reconhecido como sendo *ambiguidade da análise retórica.* Poder-se-ia dizer que essa ambigüidade retórica é anterior ao reconhecimento na ordem de ocorrência das proposições e anterior à análise retórica, pois é de conhecimento acadêmico que a própria língua natural, na sua usual concretização, é ambígua, justificando-se, portanto, tal ambigüidade.

No que se refere à definição das relações retóricas, os autores Mann e Thompson [24] identificam o Efeito como sendo um elemento essencial para auxiliar na restrição ao uso inapropriado de uma determinada relação. As definições das relações e dos esquemas só são aplicadas se o que foi apresentado pelo escritor, a fim de ativar o efeito, também fizer sentido para o analista. Nota-se que a aplicabilidade da definição de uma relação não está diretamente relacionada com a forma como a análise de um texto está sendo realizada. De forma semelhante, reforça-se que a *RST* representa as estruturas das funções e as estruturas das formas e é nesse sentido que as relações retóricas identificadas em um texto equivalem à base da coerência textual, o que é essencial para o entendimento do texto.

A estrutura do discurso é composta por proposições e a relação que se estabelece entre as proposições é definida como relação *proposicional.* Os autores afirmam que as relações não necessitam estar assinaladas na estrutura do texto por elementos de superfície, tais como marcadores ou operadores para estabelecerem relações proposicionais. As relações proposicionais podem se manifestar a partir de uma outra parte estrutural explicitada na superfície do texto. Em termos de quantidade, esse tipo de relação pode ser observado tanto quanto a realização das orações independentes. Assim, pode-se explicar as relações marcadas e as não-marcadas no texto através da parataxis e da hipotaxis, isto é, os processos que envolvem a coordenação e a subordinação.

Assim sendo, a RST apresenta uma base descritiva para estudar as relações entre as sentenças do texto em termos funcionais considerando, para tal, a distinção entre as proposições núcleo e satélite e a hierarquia entre eles. Nesse sentido, pode ser reconhecida como um *sistema* que apresenta uma combinação de características capaz de representar a estrutura hierárquica e a nuclearidade independente da amplitude do texto. A nuclearidade, na visão dos autores, é assumida como sendo *o princípio central da organização do texto*, priorizando-se o núcleo

em relação ao satélite. A afirmação é justificada a partir das evidências constatadas na investigação, na qual observou-se que se ocorrer o apagamento de um núcleo, em uma relação, o conteúdo do satélite tornar-se-á muitas vezes incompreensível, no entanto, o contrário não é verificado, isto é, caso ocorra o apagamento de um satélite, ainda é possível identificar a informação contida na estrutura a partir da informação presente no núcleo. Poder-se-ia, desta forma, reconhecer que o satélite ganha sua significação através da relação que se estabelece com a proposição nuclear.

Em termos pragmáticos, uma típica análise *RST* começa pela divisão do texto em unidades mínimas significativas, o que equivale a estruturas independentes. Cada proposição é constituída por segmentos, aos quais lhe são atribuídos um papel de núcleo ou de satélite, entre os quais pode-se atribuir uma relação retórica específica. O resultado da organização das relações reflete em uma estrutura hierarquicamente elaborada do texto, sendo que a maior extensão criada é aquela que representa o texto em sua totalidade. Ressalta-se que a maioria das relações estabelecidas entre os segmentos são assimétricas, podendo ser entre um núcleo e um satélite ou entre diferentes núcleos.

2.1.1 Revisão crítica sobre a RST

A *RST* tem sido utilizada em um grande número de aplicações em linguística computacional, devido à possibilidade de implementação em um sistema computacional. As abordagens que mais se utilizam da teoria retórica relacionam-se às aplicações que envolvem geração de texto, análise sintática, sumarização, tradução automática entre outras. No entanto, é a área da geração em língua natural - *NLG* é a que mais se utiliza das relações *RST's* ou similares, como é, em parte, o caso desta proposta.

Devido a ampla utilização que tem sido feita da *RST*, é normal que sejam apresentadas críticas sobre o alcance da teoria. Taboada and Mann [2] [42] descreveram algumas possibilidades e aplicações da RST em diferentes áreas; avaliaram os resultados práticos do uso da teoria e discutiram os problemas e dificuldades que podem ser geradas com aplicabilidade da teoria retórica. A partir das colocações apresentadas pelos autores, destacamos seguintes pontos que estão relacionados à proposta em questão:

- Identificação e Divisão das Unidades de Análise Textuais

  Conforme os autores, uma das dificuldades na aplicação da teoria retórica está relacionada à identificação e à divisão das unidades de análise, bem como, com a sua

[2] O artigo referido foi elaborado e revisado por Willian Mann e sua equipe. Willian Mann faleceu em 2004, um ano antes da edição final do artigo, a qual ficou sob a responsabilidade de Maite Taboada e colaboradores.

extensão e os seus limites. Taboada e Mann [42] discutem o assunto referindo que as unidades de maior extensão apresentam a possibilidade de estabelecer relações com elementos que se encontram fora das unidades do escopo de análise. Essa tendência, apontada pelos autores, pode ser observada no momento em que são analisados capítulos de livros, por exemplo. Conforme Marcu [29], a análise das unidades textuais de maior tamanho tendem a ser arbitrárias e podem comprometer a informatividade. No entanto, quando o nível de análise estabelece relações fora dos pares proposicionais, outros elementos, tais como os constituintes de gênero, os elementos coesivos e a própria macroestrutura são avaliados para auxiliar na identificação mais precisa das relações retóricas.

- Granularidade na Identificação das Relações Retóricas

  A granularidade no processo de identificação das relações retóricas tem sido um dos desafios mais complexos para o analista. O analista tem que estabelecer previamente quais os critérios que nortearam o processo de análise e a partir deles determinar a divisão das unidades, sua extensão e limites, tendo sempre em mente o seu objetivo final.

  Conforme os autores, Taboada e Mann [42], a situação de granularidade pode ser contornada se o processo de identificação das relações for realizado no interior das próprias estruturas oracionais, isto é, *intra-oracional*. No entanto, essa possibilidade é questionada, quando se trata das aplicações que envolvem a geração em linguagem natural, neste caso, uma baixa granularidade poderá apresentar melhores resultados.

  Carlson e Marcu [2] trabalham em busca do equilíbrio entre a granularidade e a identificação das unidades (núcleo - satélite). Os autores elegem as orações como unidades elementares do discurso – *edu's* e utilizam-se de índices lexicais e sintáticos para auxiliar na limitação das estruturas.

- A identificação das unidades de análise – proposições nucleares e satélites

  Conforme Taboada e Mann [42], no escopo da teoria retórica -*RST*, para realizarmos uma análise é necessário dividirmos o texto em unidades mínimas. A identificação das unidades é realizada antes de iniciarmos o processo de análise, para evitar qualquer tipo de circularidade, isto é, de que a análise possa estar na dependência das unidades ou de que as unidades escolhidas estejam dependendo da análise.

  Para a realização de determinadas análises, os autores postulam a existência de uma regra prática para a divisão das estruturas, ou seja, que cada oração independente e

os seus complementos constituam uma unidade. Esta possibilidade de divisão aplica-se razoavelmente a muitos objetivos, mas pode ser problemática em análises mais específicas. Neste sentido, a aplicação da regra de divisão pode apresentar problemas em relação: ao detalhamento, perdendo-se informações importantes para compreensão da unidade; à ligação da regra à linguagem do texto e ao processo de formação das orações que o compõe, caso o texto apresente uma linguagem diferente da proposta na regra e; em relação à linguagem falada em que as unidades são constituídas pela entonação não sendo necessariamente orações independentes, conforme a descrição da regra.

Definidas as unidades de análise, identificam-se os segmentos, que expressam as proposições simples entre as quais são estabelecidas relações retóricas. A proposta da *RST* é explicar texto a partir destas relações, sejam elas relações nucleares – *núcleo e satélite* ou multinucleares *núcleo e núcleo (s).* Em geral, as proposições nucleares podem ser entendidas por si próprias, contrariamente à proposição satélite. Carlson e Marcu [2] salientam que um núcleo e um satélite raramente podem ser identificados isoladamente. Em alguns casos, quando a informação semântica presente na estrutura das proposições é muito similar, pode gerar determinações nucleares diferenciadas, dependendo do contexto e de frases pistas.

Os autores sugerem dois testes para distinguir núcleos e satélites, são eles:

– teste de apagamento - quando um satélite de uma relação é apagado, o segmento mais à esquerda, isto é, o núcleo poderá ainda representar alguma função no texto, embora possa ser um pouco confuso. Caso o núcleo seja apagado, o segmento que resta, isto é, o satélite, revela-se muito menos coerente.

– teste de substituição - diferentemente do núcleo, um satélite pode ser recolocado com uma informação diferente sem alterar a função do segmento.

2.2 Aplicações da RST

Taboada e Mann [41] apresentam um relatório técnico aprofundado em que descrevem a aplicabilidade da RST em trabalhos de áreas diferenciadas. Conforme os autores, as áreas que mais fazem a utilização da RST são aquelas que desenvolvem estudos em aplicações computacionais, visto que, a RST apresenta condições que favorecem à possibilidade de implementação de sistemas computacionais. Assim, os autores reforçam a existência de trabalhos em linguística computacional, tais como: geração automática de texto, sumariza-

ção, análise automática ou *parser*, sistemas de pergunta-resposta, tradutores automáticos e avaliação de argumentos.

Na sequência deste capítulo, apresentamos importantes investigações que utilizam os princípios descritos pela RST, os quais estão diretamente relacionados à proposta metodológica desenvolvida no âmbito desta tese, devidamente testada em forma de sistema automático, denominado por AuTema-Dis.

2.2.1 O Trabalho de Daniel Marcu

Daniel Marcu é um dos autores que mais tem estudado e aplicado a RST em suas pesquisas. O autor desenvolve um estudo profundo no que se refere à sumarização automática e aos modelos estatísticos para análise discursiva. Além das investigações referidas, o autor desenvolveu um analisador discursivo de nível retórico para o Inglês, tipo *parser*, o qual pode ser aplicado em diferentes tipologias de textos. O autor é responsável por importantes avanços na RST, podemos citar como exemplo a identificação e solução de problemas relativos à parte computacional que a teoria proposta de Mann e Thompson [24] apresentava para automatização da análise retórica. Entre os aspectos elucidados por Daniel Marcu em relação à automatização das relações retóricas, observamos:

- A determinação das unidades mínimas de significação nos textos

  No que se refere às estruturas do texto entre as quais podem ser estabelecidas relações retóricas, Marcu introduz o conceito de unidades mínimas de significação ou *EDU's*, o autor afirma que a determinação das edu's está relacionada ao problema da segmentação textual. Para solucionar essa questão, utilizou duas técnicas: uma baseada na análise de corpus; outra baseada em técnicas de aprendizado de máquina.

  Na primeira técnica, Marcu Daniel [25; 26] produziu várias regras a partir da análise de corpus para a determinação dos segmentos de um texto. Para a construção de tais regras, foram considerados os sinais de pontuação e os marcadores discursivos; esses por serem indicativos superficiais da estrutura textual. O autor determinou que para cada padrão de item léxico, nesse caso os marcadores, está associada uma ação, ou seja, se identificarmos um marcador discursivo tal como *Although* no início de uma sentença, temos um item léxico - padrão lexical. Assim sendo, na sequência, deve ser realizada uma ação, isto é, deve ser inserida uma marca de fim de segmento imediatamente após a próxima ocorrência de vírgula. As ações são responsáveis por informar ao analisador retórico onde inserir as marcações de início e fim de segmento.

No caso particular do marcador *although* - dependendo de onde ocorre na sentença, pode indicar uma relação retórica - de concessão ou de contraste - entre os segmentos, em que o marcador é observado.

Daniel Marcu [25; 26] utiliza como base um estudo de *corpora* de gêneros diferenciados, assim, organizou uma lista de marcadores discursivos, suas posições no texto e quando eles causavam a segmentação no texto. Desta forma, ele identificou 11 ações que deveriam ser executadas para segmentar um texto, tendo em vista os marcadores e sua posição na estrutura. Nota-se que, o autor procurou associar a cada marcador discursivo (considerando seu contexto de ocorrência), a uma relação retórica específica. Na visão do teórico, os marcadores discursivos são elementos coesivos formados por uma ou mais palavras que explicitam o relacionamento que existe entre as partes do texto. Esses elementos, os marcadores, determinam e são determinados pela estrutura e conteúdo textual, sendo capazes de indicar a função discursiva de trechos de texto e suas contribuições para a comunicação.

Marcu opta por trabalhar apenas com os marcadores discursivos, desta forma, os diferencia dos *sentenciais* e dos *pragmáticos*. O autor justifica a escolha pelos marcadores discursivos ao demonstrar que, apesar dos marcadores pragmáticos e sentenciais possuírem formação semelhante aos discursivos, são distinguidos pelo fato de não refletirem a estrutura discursiva do texto. Ao edificar a análise, o autor utiliza padrões lexicais obtidos por meio do estudo de um corpus específico, segundo o qual, os marcadores sentenciais são utilizados para ligar as partes do texto, não apresentando função discursiva, como é o caso do *E* na estrutura - *João e Maria são irmãos*. Os marcadores pragmáticos têm por característica remeter o leitor ao seu conhecimento de mundo, conforme o exemplo *João foi preso de novo*, na frase apresentada, o leitor é remetido a relembrar a uma outra vez que João tenha sido preso em outra ocasião. O autor faz questão de demonstrar a diferença apoiando-se na função discursiva desses elementos discursivos.

Em relação às conclusões obtidas na investigação sobre a ocorrência e comportamento de alguns marcadores, Marcu observou em seu estudo que há situações discursivas específicas em que não se evidencia a presença concreta de um marcador discursivo entre os segmentos. Para tratar desse tipo de ocorrência, o autor propôs a aplicação de algumas heurísticas de simples elaboração, tornando possível inferir a relação entre os segmentos. Tais heurísticas atuam por exemplo no nível das repetições lexicais em segmentos anteriormente apresentados na estrutura textual.

No que se refere às técnicas de aprendizado de máquina, utilizadas para auxiliar na identificação das Edu's, Marcu opta pelo Classificador C4.5 (Quilan,1993), que lista

determinadas características ou *features*, as quais são responsáveis pela determinação se um item lexical aponta ou não para a presença de uma marca de segmento. Com essa técnica, o autor conseguiu um desempenho de 97% na aplicação do algoritmo. As principais características para identificar se um item lexical indica ou não a marca de um segmento são:

- A classe gramatical do item lexical sob análise;
- As classes gramaticais dos dois itens que precedem e seguem o item lexical sob investigação;
- Se o item lexical é um marcador discursivo;
- Se o item lexical é uma abreviatura;
- Se há verbos nas proximidades do item lexical sob a análise.

Considerando os resultados obtidos no processo de identificação das Edu's, isto é, dos segmentos representativos das proposições textuais, Marcu desenvolve um projeto no domínio da etiquetagem discursiva, visando a construção de um conjunto de regras a serem utilizadas em processos de segmentação. Assim, Carlson e Marcu [2] produziram um manual que, entre outras propostas, discute questões sobre segmentação textual, estruturas encaixadas nos discursos, atribuição verbal, frases-pista e relações retóricas. O manual serve como auxiliar no reconhecimento dos unidades textuais (Edu's), apresentando as possibilidades de ocorrência. Conforme a proposta do manual, é possível determinar se específicas estruturas são verdadeiras unidades de texto, passíveis de estabelecerem uma relação entre elas. Na visão dos autores, as estruturas referenciadas abaixo podem ser identificadas como Edu's:

- Orações principais;
- Orações subordinadas com marcadores discursivos;
- Complementos oracionais de verbos atributivos [3];
- Orações coordenadas;
- Orações temporais - *as que expressam o tempo em que ocorreu o evento que é apresentado na estrutura.*
- Orações relativas.

Na investigação realizada, os autores determinam que algumas estruturas não devem ser identificadas como segmentos, tais como, as orações que desempenham as funções de sujeito ou objetos dos verbos; e os complementos verbais oracionais.

[3] Verbos atributivos: são verbos que atribuem uma expressão a algo ou alguém. Essa expressão pode se manifestar como uma fala ou um pensamento.

- A determinação das relações retóricas intra e intersentenciais

  No que ser refere às relações retóricas, é importante definirmos que a retórica é a forma como são expressas as intenções do produtor em discurso. Através das relações retóricas, um determinado autor organiza e estrutura o conteúdo informacional, para que o seu objetivo seja cumprido, isto é, que um leitor ou um ouvinte consiga reconhecer a proposta do produtor a partir de uma estrutura específica. Observa-se que existem muitas maneiras em se fazer manifestar uma mesma intenção, consequentemente, há muitas estratégias retóricas para tal e vice-versa. Essa possibilidade dificulta a identificação e a classificação das relações retóricas, o que pode ser observado na literatura como a ambigüidade na análise retórica, conforme apresentou Mann e Thompson [24].

Muitos autores estudam as relações retóricas, no entanto, Daniel Marcu [25; 27] aprofunda a teoria e expande o conjunto das relações propostas por Mann e Thompson [24] de 25 para 78 relações de ordem semântica, organizadas em 16 classes específicas. Neste conjunto, o autor identificou além das relações semânticas, relações de cunho estrutural. Na definição dessas relações, Marcu determina como característica principal o fato delas não apresentarem um significado aparente, mas auxiliarem na estruturação retórica de textos. Estas relações são de dois tipos –*Parenthetical*, as quais apresentam uma informação acessória, extra, não-linear no texto, como por exemplo: as informações entre parênteses, colchetes, chaves, em notas de rodapé ou notas explicativas; e a relação – *Same-Unit* é utilizada para unir segmentos textuais não-adjacentes que expressam uma única proposição, como por exemplo, as relações relativas que ocorrem no interior de proposições.

Em outro trabalho, Marcu e Echihabi [28] utilizaram um classificador Naive-Bayes para determinar as relações retóricas entre duas proposições textuais. Os autores elegeram como características – *features*– as próprias palavras das proposições para auxiliar a tarefa de aprendizado. Os autores elegeram as relações retóricas de contraste, evidência, explicação, condição e elaboração para testarem o classificador, mesmo quando não estavam presentes frases-pistas. Com esse procedimento, os autores pretendiam capturar dois tipos de conhecimentos:

- que marcadores discursivos indicam relações retóricas;
- que tipo de conhecimento mundo é representado por tais relações.

Neste caso da representação do conhecimento de mundo, os autores demonstram que o classificador aprende as características necessárias para identificar que tipo de relação pode ser estabelecida a partir das características das próprias palavras que compõem as proposições entre as quais se estabelece a relação retórica. Os autores apresentam

como exemplo os pares de palavras *(good, fails)* e *(embargo, legally)* como bons exemplos para indicar uma relação de contraste.

Em 2003, Soricut e Marcu [40] utilizaram um modelo probabilístico para determinar as relações retóricas intra-sentenciais utilizando como base a informação sintática e lexical. Para realizarem o estudo, utilizaram textos do estilo jornalístico. Soricut e Marcu propuseram modelos distintos para realizar a segmentação textual e detectar as relações retóricas.

No que se refere ao modelo probabilístico para segmentação textual, os autores elaboraram um modelo que utiliza as palavras (núcleos – heads) dos constituintes (sujeitos, objetos e predicativos) das estruturas sintáticas lexicalizadas. O modelo é treinado com palavras específicas e a cada nova realização de uma específica palavra poderá indicar, por critérios probabilísticos, um novo segmento. No que se refere à identificação das relações retóricas, os autores utilizam-se dos núcleos (heads) contidos nos segmentos das proposições classificadas com as relações retóricas, para determinar nas novas sentenças as novas relações retóricas, tendo sempre como base os *núcleos*.

- A determinação dos segmentos Núcleos e dos segmentos Satélites

  As relações retóricas são estabelecidas entre segmentos simples do texto. Para determinar quais segmentos são Núcleos e quais são os Satélites na estrutura, Marcu [25; 26] observa a ordem preferencial dos núcleos e dos satélites de cada uma das relações retóricas e avalia quais proposições podem funcionar como núcleo e como satélite. Para tal estudo, o autor recorreu à análise de corpus realizada e associou aos marcadores discursivos compilados às possíveis ordenações dos segmentos. O autor optou novamente pelo estudo do marcador *although*, observando os casos em que o marcador aparecia em posição inicial na estrutura. Marcu constatou com a análise que a proposição a qual o marcador pertence representa um satélite, e a proposição expressa pelo segmento posterior representa o núcleo.

- A elaboração de estruturas retóricas válidas

  Para Marcu [25], o principal desafio foi a elaboração de estruturas retóricas válidas a partir das relações retóricas que se estabelecem entre os segmentos do texto. O autor afirma que tal dificuldade resultava da falta de formalização que não permitia que se automatizasse essa etapa. Observado esse fato, Marcu procedeu a uma completa formalização da *RST* o que resultou na possibilidade da automatização da análise retórica. Os principais pontos desenvolvidos pelo autor e que permitiram a completa automatização foram:

- Critério da composicionalidade - em que são aplicadas regras que recursivamente chegam aos nós-folhas das estruturas a serem combinadas.

- Com base no critério de composicionalidade, considerando um conjunto de relações retóricas estabelecidas entre os segmentos textuais, torna-se possível montar várias estruturas retóricas válidas para um mesmo texto. Para montar apenas estruturas retóricas válidas, o autor desenvolveu uma abordagem logarítmica. O algoritmo mais utilizado e que apresenta melhores resultados é o *DCG - define clause grammar*. Para tal, considerando-se um dado conjunto de relações retóricas entre os segmentos, produz-se uma gramática em *DCG* que gera todas as combinações possíveis entre as relações, elaborando, assim, as estruturas retóricas válidas.

Além dos esclarecimentos sobre a *RST*, Marcu desenvolveu outros trabalhos significativos no que se refere à segmentação textual, identificação e classificação das estruturas retóricas, os quais eventualmente serão referidos no âmbito desta tese.

Pela seriedade no desenvolvimento da pesquisas desenvolvidas pelo autor, pode-se constatar que os resultados por ele apresentados são relevantes aos trabalhos que envolvem o estudo e a aplicabilidade das relações retóricas em diferentes áreas de investigação. No entanto, como apresentamos no início desta seção, sua contribuição relaciona-se especificamente às áreas da sumarização automática, tradutores e modelos estatísticos para análise discursiva, *parser* retórico, modelos probabilísticos e ferramentas de anotação.

Em nossa investigação, a contribuição do trabalho de Marcu está direcionada ao uso da ferramenta para catalogar as relações retóricas do corpus analisado; a utilização dos critérios estabelecidos pelo autor para determinar se uma específica ocorrência linguística pode ou não ser considerada um segmento em uma estrutura discursiva e o detalhamento proposto para segmentação e etiquetagem discursiva.

2.2.2 Trabalho de Michael O'Donnell

O autor Michael O'Donnell desenvolve uma ampla investigação relacionada ao tratamento computacional de textos e em linguística sistêmico-funcional. No que se refere às linhas de pesquisa, a sua abordagem compreende a modelagem do discurso, a geração de texto e sentença, ferramentas para realizar análises em textos, sumarização de documentos, sistemas de avaliação, representação do conhecimento e dos dados conceituais e linguísticos, a interação

estrutural e formalismo sistêmico (formalismos para modelar linguagem, comportamento e conhecimento). Em seus primeiros trabalhos, no período de 1990 até 1999, Michael O'Donnell dedica-se ao estudo de estrutura do diálogo, a modelagem da interação discursiva da fala e modelagem da fala em sistemas de telecomunicação.

No entanto, em seu trabalho de *doutoramento*, Michael O'Donnell [33] direciona-se à análise e à geração de sentenças, desenvolve um sistema computacional para análise e geração de sentenças usando um Sistema Linguístico-Funcional– (SFL). O sistema é estruturado em duas partes: a primeira trata do recurso sistêmico usado para representações linguísticas, em que a sentença é modelada em termos semânticos, léxico-gramatical e grafológico. A segunda parte descreve o processo que usa esse recurso *(recurso sistêmico)*, focalizando em uma única sentença a análise e a geração, trata-se de um estudo voltado à complexidade e à estruturação de sentenças.

No que se refere às sentenças e aos recursos textuais presentes em um texto, O'Donnell apresenta considerações relevantes a esse respeito. Para O'Donnell [33], a frase é uma unidade *grafológica* definida em termos de palavras e marcas de pontuação, a qual tipicamente representa uma estrutura gramatical ou estrutura [4] complexa. Há, no entanto, uma típica correspondência entre as frases e as unidades neste nível, pois as frases normalmente expressam atos de fala isolados, tais como: questões, ordens, comandos e afirmações. Na mesma ordem, O'Donnell utiliza a expressão enunciado no lugar de frase. No caso do processamento de frases, O'Donnell parte de uma representação abstrata e, a partir dela, produz uma frase que a expressa - *representa*. Neste sentido, o processamento pode ser identificado como uma forma variada de representações de uma frase.

Em relação aos *Recursos Textuais*, O'Donnell [33] divide os recursos de linguagem em dois processos distintos: *micro e macrorrecurso textual*. Os microrrecursos estão relacionados às unidades linguísticas e são co-extensivas com a sentença (sentenças, palavras e caracteres); e os macrorrecursos relacionam-se às representações multi-sentenciais do texto tanto no estrato grafológico (parágrafos, seções) quanto no estrato semântico (interação ideacional e estruturação textual multisentencial). A partir dessas definições e caracterizações, o autor mostra como um texto multisentencial é organizado para realizar a macroestrutura textual, tendo em vista uma específica mensagem.

Considerando a perspectiva apresentada por O'Donnell, um texto é analisado a partir dois níveis específicos: o nível dos macrorrecursos (recursos multisentenciais) e nível dos microrrecursos (recursos de uma única sentença). Conforme tal abordagem, o texto é visto sob uma

[4]Clause: a palavra é identificada nesta tese como sinônima de frase, podendo ser considerada, em alguns exemplos, como oração, conforme a definição proposta por O'Donnell [33] página 23.

perspectiva de representação interacional, isto é, como uma das partes que estão envolvidas no processo interativo entre participantes de um evento social. A análise de tais recursos e níveis é realizada, conforme O'Donnell [33], em três dimensões distintas para cada nível:

- Os macrorrecursos:
  Estrutura Temática – como um texto é estruturado para desenvolver um tema ou temas;
  Estrutura Retórica – como um texto é organizado para encontrar os objetivos retóricos do falante;
  Status da Informação – como a informação apresentada em um texto pode ser recuperada e identificada.

- Os microrrecursos:
  Tematicidade – relaciona-se à estrutura temática, informação central;
  Relevância – relaciona-se à estrutura retórica;
  Recuperabilidade e identificabilidade – relaciona-se à estrutura da informação.

O'Donnell [33] avalia a tematicidade ou *tema* considerando dois níveis estruturais distintos, são eles:

- a estrutura macrotemática – refere-se à organização ou modelagem do tema, isto é, o meio como a mensagem, contida em um texto, é estruturada. Na visão do autor, um texto é organizado de tal forma para que possibilite a um leitor ou a um ouvinte identificar o que pretende o autor do texto com aquela estrutura;
- a estrutura microtemática – refere-se ao tema de uma sentença, isto é, o *ponto de partida* para a composição da mensagem como um todo. O autor ressalta que definição do tema está associado ao que propõe Halliday (1985), ou seja, a noção de *topical thema*.

Em relação à RST, O'Donnell identifica a estrutura textual em termos de relações de dependência entre as unidades dos textos. Ao desenvolver um estudo a respeito do processo da geração de texto, o autor considera a relevância retórica como um dos pontos a serem considerados no processo. A seleção do conteúdo, isto é, a seleção da informação é a base para a apresentação do conhecimento de um falante. No nível da frase, o autor identifica a relevância para determinar quais os papéis que podem ser expressados por uma proposição, quais os participantes desse processo são realmente importantes.

A representação da relevância é um processo dinâmico, variando de acordo com a progressão textual. O'Donnell usa a noção de relevância, noção essa derivada da relevância em *geração de texto*, em que a informação é selecionada como relevante pelo falante para desenvolver os objetivos do discurso. A noção de espaço relevante está associada ao conjunto dos processos ideais, participantes e circunstâncias que são relevantes no que concerne aos objetivos do discurso.

2.2.2.1 A RST no trabalho de O'Donnell

As Relações Retóricas, propostas por Mann e Thompson [24], são referidas por O'Donnell [33] como sendo relações de dependência que se estruturam entre as unidades de um texto. Elas são usadas no processo que constitui a geração de texto e têm como finalidade a construção da coerência de um texto. A RST identifica os segmentos presentes em um texto e as *relações retóricas* entre eles. O'Donnell enfatiza que, enquanto a RST pode ser utilizada para análise do discurso por seres humanos, no que se refere à análise computacional do discurso, ela apresenta algumas restrições devido ao fato que muitas das relações retóricas não são marcadas explicitamente no texto ou se apresentam de forma ambígua, caso que impossibilita, algumas vezes, recuperá-las na análise do texto.

O problema das relações retóricas não marcadas evidenciadas por O'Donnell, foi também observado por Daniel Marcu [25; 26] ao estudar os marcadores discursivos como indicadores de segmentação. Para resolver essa dificuldade, Marcu desenvolveu heurísticas simples que auxiliam na identificação dos segmentos que representam as proposições do discurso e, consequentemente, auxiliam na identificação das relações entre tais segmentos.

Apesar de não ter apresentado uma solução para o problema das relações não marcadas, O'Donnell desenvolveu um dos mais significativos trabalhos [34; 37], a ferramenta *RSTTool* – para a marcação retórica de corpus, a qual utiliza-se basicamente das relações retóricas propostas por Mann e Thompson [24].

A ferramenta desenvolvida por O'Donnell dispõe de uma interface gráfica que facilita a marcação das relações retóricas na estrutura do texto, bem como, é possível a sua utilização para realizar o processo de segmentação. A ferramenta é eficaz e auxilia a estruturar textos, possibilitando reformular quaisquer das ações realizadas durante o processo de anotação, além disso, a ligação gráfica da ferramenta gera o texto segmentado em uma estrutura arbórea.

Observa-se que a ferramenta *RST-Tool* pode ser utilizada em dois momentos específicos, isto é, quando se quer marcar os limites dos segmentos e quando desejamos ligar os segmentos graficamente, gerando as árvores estruturais. O uso da ferramenta tem como vantagem: a redução no tempo de análise, a qual permite criar e modificar estruturas, quando necessário; a facilidade na preparação das figuras, as quais descrevem as relações retóricas em árvores. Além disso, a *RST–Tool* provou ser útil na sua função de geração de texto, na apresentação de documentos de extensão variável. Além da sua aplicação principal, poder ser utilizada como ferramenta nos processos que realizam sumarização, apesar de não ser essa sua aplicabilidade inicial. Como o nome diz, é uma ferramenta à disposição do usuário no sentido de estruturar a organização arbórea, a ferramenta não atribui automaticamente as relações retóricas, estas ficam à escolha dos usuários da ferramenta.

Entre outros trabalhos, O'Donnell [35] desenvolveu um estudo na área de sumarização, trata-se de um sistema que é capaz de identificar os segmentos textuais mais relevantes para comporem um sumário. O funcionamento do sistema está na dependência da marcação dos documentos com a *RST- Markup Tool*, que é utilizada para *podar* o texto conforme o tamanho requisitado pelo usuário, selecionando apenas o que é essencial do texto. No processo de seleção e poda do texto, o usuário da ferramenta tem total liberdade para ajustá-la conforme a sua necessidade ou demanda. O maior problema apresentado por esse sistema está associado à restauração da coerência depois do texto *podado*, em especial, com os elementos referenciais, os marcadores discursivos, a pontuação e a organização dos parágrafos. Além do problema com a restauração da coerência, O'Donnell aponta para a questão que diz respeito à nuclearidade, isto é, quando o conteúdo mais relevante não se encontra em posição nuclear.

Em 1997, Knott e O'Donnell [19] dedicam-se à elaboração de regras restritivas, as quais são empregadas no sistema de geração de texto *ILEX-I*. Trata-se de um conjunto de regras que restringem as possibilidades de seleção do conteúdo e estruturação dos componentes do discurso. Para tal, descrevem uma arquitetura que incorpora regras para a modelagem do discurso.

Como conclusão, os autores ressaltam que a representação de regras restritivas no *ILEX-I* servem como uma orientação para agrupar fatos isolados de um domínio, além de permitirem que o usuário e o sistema discutam as conceitualizações de domínio. Na visão dos autores, a utilização das regras no sistema são essenciais para:

1. realizar os objetivos educacionais do sistema;
2. realizar a coerência do texto através do uso de variadas e interessantes relações de

coerência;

3. endereçar as concepções que o usuário talvez tenha avançado ou talvez tenha desenvolvido durante a sessão.

Na sequência das investigações, O'Donnell [36] desenvolve um estudo relacionado à composição automática de texto relacionando múltiplas estratégias discursivas requeridas pelo contexto de escrita. O autor propôs uma abordagem para compor textos automaticamente, considerando as diferentes estratégias utilizadas pelos produtores ao elaborarem diferentes tipos de textos. O'Donnell afirma que o estímulo para desenvolver esse trabalho surgiu a partir dos resultados obtidos no *ILEX-I*[5], o qual contém conhecimento de como expressar os fatos unidos por várias relações retóricas – *RST's*. Como conclusão, o autor identifica que a produção de um texto está relacionada às múltiplas estratégias de produção discursiva, sendo possível aplicá-las à composição automática de textos.

No ILEX I, o autor observa a necessidade que a geração de texto em hipertexto tem em lidar com ambientes não-interativos. Nesse sentido, propõem uma arquitetura para estruturar esses diferentes ambientes, os quais são descritos em níveis de uso da representação linguística e os processos que se estabelecem entre eles. Nesse sistema, a RST é utilizada para organizar o conteúdo e gerar as árvores de dependência. Conforme o autor, as relações retóricas que existem entre todos os *nós-texto* e correspondem aos nós das relações que ligam os fatos em um índice potencial. Assim, o projeto destina-se:

1. À edificação e avaliação de três gerações de um sistema de trabalho, sendo cada uso demonstrável em uma aplicação real;

2. Ao desenvolvimento de teorias na área da estrutura do discurso e de algoritmos de geração de discurso;

3. À apresentação e avaliação de novos tipos de linguagens materiais do mundo real em dois domínios da ficção (rótulos de exibição nos museus e catálogos de compras domiciliar).

Em 2003, Michael O'Donnell e outros investigadores [32] revisam um de seus trabalhos em análise discursiva e apresentam aspectos relevantes relacionados ao analisador. O autor afirma ser necessário utilizar uma gramática funcional para dar suporte ao analisador, sendo o *parsing grammar* um sistema funcional, que amplia o domino da gramática.

[5]ILEX:Intelligent Labelling Explorer

2.2.3 Estudos Desenvolvidos no NILC - Núcleo Interinstitucional de Linguística Computacional

O NILC – *Núcleo Interinstitucional de Linguística Computacional* - da Universidade de São Paulo, Brasil – desenvolve projetos de pesquisa em linguística computacional voltados ao processamento da linguagem natural. No que se refere aos interesses do grupo, observam-se pesquisas para o desenvolvimento do Corpus e do Léxico, investigações a respeito dos processos de Sumarização Automática, Tradução de Máquina, além do desenvolvimento de Ferramentas Computacionais que fornecem suporte à escrita.

O grupo de investigação do *NILC* é composto por pesquisadores de diferentes áreas, fato que imprime ao grupo uma característica interdisciplinar. Neste sentido, o *NILC* abarca diferentes projetos de pesquisa, entre eles, os projetos: DIZER, Rhetalho e SENTER; o RHeSummaRST; o EXPLOSA e o DMSumm que se interrelacionam e se complementam. No entanto, nosso interesse pelos trabalhos desenvolvidos no âmbito do NILC está diretamente relacionado aos projetos que investigam a análise discursiva e utilizam a RST como uma ferramenta para auxiliar no processo estrutural do discurso.

2.2.3.1 Projetos NILC relacionados à RST

- O projeto DiZer *DIscourse analyZER* foi desenvolvido como parte do trabalho de doutoramento de Thiago Alexandre Pardo [38] e orientado por Maria das Graças Volpe Nunes. Trata-se de um analisador discursivo automático para a língua portuguesa do Brasil. No que se refere à estrutura do analisador *DiZer*, a equipe de investigação elaborou um modelo e um protótipo para realizar análise retórica automática, extraindo de um texto a sua estrutura profunda, isto é, realizar a construção automaticamente das estruturas discursivas. Como diferencial na composição do analisador, o autor ressalta a utilização de modelos estatísticos inéditos baseados em unidades de conteúdo de crescente complexidade, unidades que por seu domínio abordam palavras, conceitos e estruturas argumentais.

  Para o cumprimento de tal propósito, foram identificados os marcadores discursivos que constituem, na visão do autor, o principal mecanismo linguístico para a detecção de relações retóricas; palavras e frases indicativas relacionadas ao gênero e domínio textual. Além destes casos analisados, Pardo realizou estatísticas sobre a organização discursiva; observou o relacionamento entre palavras e conceitos para a realização da análise discursiva; verificou o relacionamento entre as estruturas argumentais subjacentes aos segmentos textuais para automatização da proposta.

O analisador discursivo *DIZER* foi construído a partir do resultado de diferentes estágios de investigação, conforme citamos:

1. o processo que envolve a segmentação automática do texto em unidades menores, sendo estas unidades do tipo: cláusulas, orações, parágrafos ou tópicos. Na etapa de segmentação, existem fatores que devem ser considerados, entre eles, a granularidade, a qual está diretamente relacionada ao tipo de aplicação que se quer dar à segmentação realizada. Segundo Pardo [38], a segmentação é uma tarefa delicada e envolve alguns problemas de difícil resolução como: os sinais de pontuação; a possibilidade de confundirmos marcadores discursivos com os outros tipos, isto é, sentenciais e os pragmáticos; e as referências anafóricas podem apresentar alguns problemas, quando se tratar de um segmentação topical.

2. a identificação das principais teorias discursivas utilizadas em PLN, considerando os níveis retórico, semântico e intencional. Para o autor, uma determinada estrutura discursiva é necessariamente formada pela relação desses três níveis. O nível retórico é reconhecido através das *relações retóricas* apresentadas pela RST Mann e Thompson [24]; o nível semântico é recoberto pela teoria de Jordan [17]; já o nível intencional fica amparado pelas relações intencionais de GSDT Grosz e Sidner [14] – *Grosz and Sidner - Discourse Model.*

3. a composição de um corpus em Português Brasileiro com marcação retórica RST. Para realizar a marcação retórica dos textos, Pardo [38] utilizou a ferramenta desenvolvida por Daniel Marcu – *RST Annotation Tool.* A ferramenta utilizada permitiu segmentar os textos, escolher as relações retóricas e identificar os segmentos núcleos e satélites. Além disso, a utilização da *–RST Annotation Tool–* permitiu que fossem armazenados todos os passos executados e alternar entre a marcação e segmentação. No processo de marcação retórica foi utilizado inicialmente o elenco das relações propostas por Mann e Thompson [24], no entanto, essas mostraram-se insuficientes para dar cobertura ao corpus. A fim de solucionar o problema, foram consideradas também as relações propostas por Marcu [25]. Como estratégia de anotação, forma adotados os seguintes procedimentos: – todas as proposições presentes em uma sentença foram relacionadas retoricamente; – na sequência, todas as sentenças de um parágrafo foram relacionadas; por fim, os parágrafos foram relacionados.

4. a análise dos *corpora* de textos científicos anotados retoricamente relacionando a ocorrência de um marcador discursivo a um tipo relação retórica. O autor

especifica em uma tabela a distribuição dos marcadores discursivos em função das relações por eles assinaladas. Na construção do *DiZer*, os resultados extraídos do conhecimento do corpus analisado foi codificado em padrões de análise e heurísticas. Tal conhecimento pode ser aplicado a novos textos, permitindo também identificar as relações retóricas e construir novamente uma estrutura discursiva.

O *DIZER* relaciona-se diretamente a dois projetos adjuntos: o *Rhetalho* e o *Senter*.

- O Projeto *Rhetalho* está relacionado ao *DiZer* e ao Projeto *RHeSumaRST*, trata-se de um corpus anotado retoricamente com as relações apresentadas na *Rhetorical Structure Theory – RST*, disponibilizado à comunidade científica. O projeto em questão produziu um corpus de referência devidamente anotado por dois especialistas em *RST*. A anotação foi realizada com o auxílio da ferramenta RSTTool com um protocolo específico para a anotação. O *Rhetalho* é composto por 50 textos: 20 secções de introdução e 10 secções de conclusão de artigos científicos do domínio da Computação; 20 textos do jornal on-line Folha de São Paulo, mais especificamente, 7 textos da Secção Cotidiano, 7 da Secção Mundo e 6 da Secção Ciência. Na perspectiva dos autores, o *Rhetalho* pode ser aplicado em dois segmentos de pesquisas diferenciados, ou seja, na linguística, no âmbito do discurso, no que se refere à análise de marcadores superficiais do discurso, anáforas e elipses; na linguística computacional, o *Rhetalho* pode ser utilizado para o desenvolvimento de sistemas que manipulam conhecimentos discursivos, tais como analisadores e sumarizadores automáticos.

- O Projeto *SENTER SENtence spliTER* tem como objetivo desenvolver um segmentador sentencial automático destinado aos textos escritos em Português do Brasil. A proposta do *SENTER* é segmentar um texto produzido em Português ou em Inglês em sentenças, utilizando-se de sinais de pontuação como elemento referencial. O segmentador apresenta considerações a respeito das URLs e e-mails, reticências, citações, parênteses e números reais entre outros sinais. O *SENTER* foi elaborado para realizar segmentação sentencial a partir de regras específicas, são elas:

  1. uma sentença é delimitada sempre que houver marca de nova linha *carriage return e line feed* independentemente de um sinal de fim de sentença ter sido encontrado anteriormente;

  2. não são delimitadas sentenças dentro de aspas, parênteses, chaves e colchetes;

3. uma sentença é delimitada quando são encontrados os sinais de pontuação: interrogação (?) e exclamação (!);

4. uma sentença é delimitada quando o símbolo de ponto (.) é encontrado e este não é um ponto de número decimal, não pertence a um símbolo de reticências (...), não faz parte de endereços de e-mail e páginas da Internet e não é o ponto que segue uma abreviatura;

5. uma sentença é delimitada quando uma letra maiúscula é encontrada após o sinal de reticências ou de fecha-aspas.

- Outro projeto que se insere no âmbito do **NILC** é o *RHeSumaRST*. O Projeto *RHeSumaRST* tem como objetivo construir um conjunto de heurísticas para textos com base nas suas estruturas retóricas – *RST*, focalizando-se especificamente nas ligações–cadeias referenciais que estão presentes nos textos-fonte. As heurísticas desenvolvidas têm como princípio reconhecer, em uma estrutura RST do texto-fonte, as subestruturas mais relevantes para a construção de um sumário. A proposta apoia-se na coerência do texto além de poder contar fortemente com a *RST* e a *Veins Theory*, sendo ambas consideradas fundamentais no processo para a construção de uma estrutura sumarizada.

  O RHeSumaRST está estruturado a partir da configuração apresentada pela *RST*, possibilitando que as relações de significado entre as estruturas do discurso possam ser recuperadas. Considerando a estrutura *RST* e a *Veins Theory*, as heurísticas formuladas delimitam o domínio de acessibilidade referencial para cada unidade do discurso, determinando os limites nos quais os antecedentes de uma anáfora podem aparecer no discurso. Como mencionado, o modelo consiste em um conjunto de heurísticas, focalizando as relações retóricas, as que representam as informações menos relevantes para serem candidatas ao processo de exclusão. A organização do sistema foi estruturada de forma que o modelo proposto possa preservar a coerência do sumário apresentado.

- O Projeto Explosa - *métodos para sumarização automática* – é outro projeto que se encontra no âmbito do *NILC*. O *Explosa*, como o nome define, explora métodos diferenciados para a sumarização automática e como diversificados focos podem participar na construção de um sumarizador automático para textos escritos em Português Brasileiro. O desenvolvimento do Explosa contou com a colaboração de vários pequenos projetos, como: o *NeuralSumm* – NEURAL network for SUMMarization ; o *GistSumm* - Gist SUMMarizer ; o *DMSumm* - Discourse Modeling SUMMarizer.

Apresentamos, sucintamente, os projetos que estão salvaguardados pelo *NILC*, os quais desenvolveram pesquisas correlacionadas ao projeto de doutoramento em questão. Nos próximos capítulos, alguns destes projetos serão referenciados novamente, devido à proximidade metodológica no âmbito deste estudo, em relação à estrutura de uma arquitetura para automatização da análise discursiva.

2.3 Ferramentas – Analisadores Discursivos Automáticos e Marcadores Retóricos –

Na literatura da área de Linguística Computacional, especificamente, na sub-área *PLN* que envolve a análise discursiva automática, observa-se o desenvolvimento de significativos trabalhos, sendo, na maioria direcionados à Língua Inglesa. A construção de analisadores automáticos discursivos vem sendo muito aprimorada, apresentado resultados significativos, conforme mostram os resultados de algumas pesquisas apresentadas nesta tese. Os sistemas analisados apresentam cada vez mais refino e acuro, pois desde a edificação da arquitetura delimita-se uma finalidade específica em conformidade com tipo de abordagem e limite para realização do sistema.

Assim, apresentamos, brevemente, as ferramentas, os analisadores discursivos e marcadores retóricos mais reconhecidos e relevantes identificados na literatura, ressaltando que apenas um dos estudos refere-se especificamente ao Português Brasileiro, conforme descrevemos abaixo:

- Marcu [25; 27] - o autor desenvolveu um parser retórico para a língua inglesa, para realizar análise de textos jornalísticos, utilizando os princípios apresentados na *RST*. O trabalho de Marcu apresenta-se em destaque, pois o autor revisou e tratou de alguns problemas da *RST*, tais como a construção de estruturas retóricas válidas considerando as relações entre os segmentos; a determinação dos segmentos nucleares e satélites; a implementação do critério de composicionalidade; a determinação dos segmentos que expressam proposições simples, e reconhecimento automático das relações retóricas intra e intersentenciais.

- Marcu e Echihabi [28] – desenvolveram um *classificador* para reconhecer relações retóricas entre os segmentos de um texto. O *classificador* trabalha com técnicas de aprendizado que utilizam como características as próprias palavras do segmento que expressam as proposições que se encontram relacionadas ao processamento em questão.

O objetivo do classificador é: identificar que relações retóricas são indicadas pelos marcadores discursivos e o conhecimento de mundo, conforme descrito na seção 2.2.1 desta tese. As relações retóricas identificadas pelo *classificador* são de contraste, explicação-evidência, circunstância e elaboração e o resultado na identificação é (93%), mesmo quando as relações não estão marcadas explicitamente no texto por frases-pistas.

- Soricut e Marcu [40] – os autores elaboraram um analisador discursivo baseado em modelos probabilísticos, os quais recorrem às informações sintáticas e lexicais. A análise é realizada intra-sentencialmente em textos de caráter jornalístico, segmentando-os e detectando as relações retóricas ao longo da estrutura. Os autores treinam um modelo probabilístico para determinar a estrutura retórica entre os segmentos das novas estruturas com base em seus núcleos, conforme descrito na seção 2.2.1 desta tese.

- Corston-Olivier [3] – o autor desenvolve um sistema computacional - *RASTA (Rhetorical Structure Theory Analyzer)*, o qual representa a estrutura de textos escritos de cunho enciclopédico. Para tal, conta com informações sintáticas lexicalizadas das sentenças do texto, além dos marcadores discursivos e de alguns aspectos da forma lógica para determinar as relações retóricas. O autor desenvolve o conceito de subespecificação, ou seja, quando não há informação suficiente para detectar a relação, o analisador indica que a relação existe, mas não especifica concretamente qual a relação.

- Thiago Pardo [38] – o autor desenvolve o analisador discursivo DiZer – *DIscourse analyZER*. Trata-se de um analisador discursivo automático desenvolvido para o Português Brasil. O analisador realiza a análise retórica automática, para tal, considera as informações linguísticas provenientes dos marcadores discursivos, palavras e frases-pistas. Para a compilação do conhecimento, o autor utilizou 100 textos de cunho científico da computação retoricamente anotados. Os processos que podem ser realizados pelo *DiZer* são: segmentação textual; detecção das relações retóricas; construção de estruturas retóricas válidas. O DiZer apresentou resultados satisfatórios para os textos científicos e para os jornalísticos que compuseram os *corpora*.

2.4 Resumo do Capítulo

Neste capítulo, revisamos os trabalhos relacionados à análise automática de discurso e descrevemos os principais trabalhos na área de *PLN*; apresentamos *RST* - Rhetorical Structure Theory - Teoria da Estrutura Retórica e a sua aplicabilidade em estudos que procuram explicar questões relacionadas à composição discursiva.

No próximo capítulo, apresentamos o enquadramento teórico-linguístico que sustenta a proposta metodológica para análise automática de discurso.

Capítulo 3

Enquadramento Linguístico

3.1 Introdução

Neste capítulo, apresentamos uma revisão linguística dos conceitos que sustentam formalmente à metodologia desenvolvida no âmbito desta tese, a qual se destina a fornecer subsídios à análise automática de discurso.

Para realizar a investigação proposta, foi necessário elaborar um constructo teórico que articulasse informações formais e conceituais relativos à produção e à recepção de um texto/discurso. Assim, desenvolveu-se uma base metodologia para a automatização da análise discursiva, a qual pudesse, ao final do processamento, apresentar uma macroestrutura/macroproposição representativa do texto analisado.

O texto é o objeto de análise desta investigação tanto na sua estrutura superficial, em que são analisadas as relações coesivas; quanto na sua estrutura profunda, em que são identificadas as relações conceituais e significativas, que remetem ao discurso implícito nele contido. Assim, para a composição da metodologia e realização automática da análise foi necessário considerar dois níveis: o nível formal relacionado à estrutura superficial (coesão) intra e intersentencial e o nível conceitual relacionado à estrutura profunda (coerência), suportado pelas relações retóricas que se articulam e organizam o tema discursivo.

3.2 Componentes Linguísticos da Metodologia

Na presente seção, apresentamos os conceitos linguísticos que suportam a base metodológica desenvolvida nesta investigação. As especificações apresentadas são relevantes e pertinentes ao estudo realizado entre elas, faz-se necessário clarificar o que entendemos por texto e por discurso; o que identificamos como segmentos e subsegmentos em uma estrutura textual; a determinação do significado de proposição e; a determinação do que é e como se constrói no âmbito desta proposta a macroestrutura/macroproposição.

3.2.1 Texto - uma concepção

Na busca por uma explicação para o uso e organização da linguagem, seja ela escrita ou falada, muitas teorias e propostas vêm sendo desenvolvidas, apoiadas em fundamentos linguísticos e filosóficos, os quais observam diferentes aspectos da linguagem, procurando apresentar análises mais elucidativas sobre a manifestação do fenômeno.

Neste sentido, desenvolver um estudo que contemple a realização de uma análise automática discursiva requer alguns esclarecimentos. Um ponto importante a ser explicitado é a definição do que entendemos por texto no escopo deste trabalho. É consensual, em estudos linguísticos, a dificuldade em conceituá-lo, bem como, apresentar as características que o diferenciam em relação ao discurso.

Costa Val [4], na avaliação que faz a respeito da definição para texto/discurso, em *Redação e Textualidade*, identifica texto ou discurso como uma *ocorrência linguística falada ou escrita de qualquer extensão, dotada de unidade sociocomunicativa, semântica e formal*, conforme mostra a figura 3.1.

Na perspectiva da autora, texto e discurso equivalem-se enquanto organização global, sendo identificada a partir da interação de unidades em que aspectos sociais, formais e conceituais relacionam-se para constituir toda a estrutura. Especificamente em relação à metodologia que propomos para análise automática do discurso, consideramos essa interação e articulação das unidades conforme apresenta Costa Val [4].

No entanto, identificar um *texto* como uma unidade de linguagem em uso, cumprindo uma função identificável num dado jogo de ação sociocomunicativa, conforme propõe Costa Val [4] não é suficiente para a metodologia que propomos. A abordagem desenvolvida para

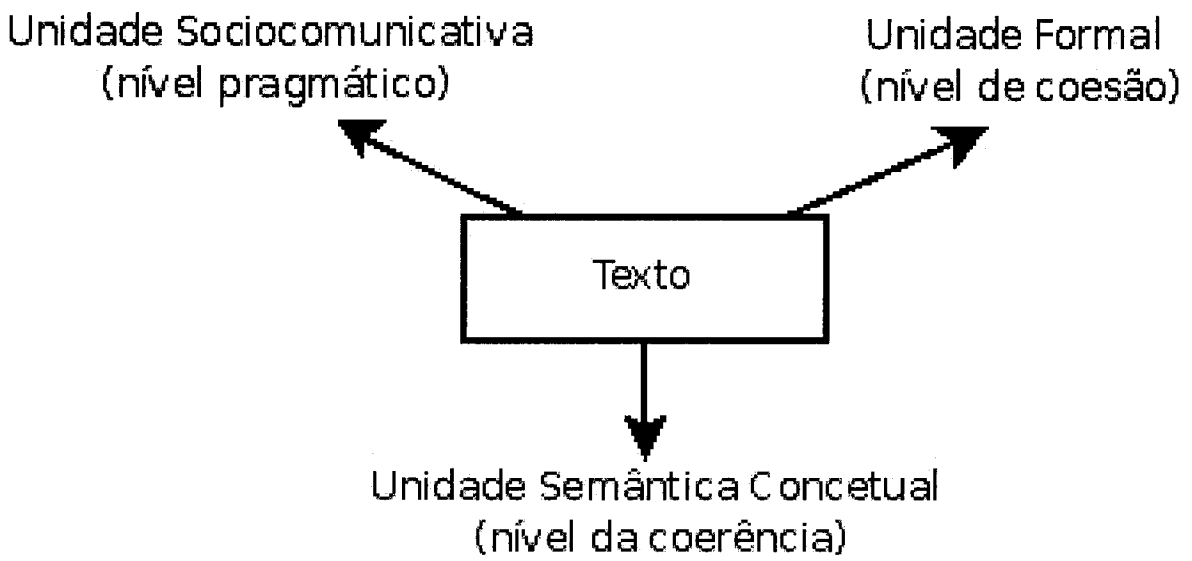

Figura 3.1: Texto: Unidade de linguagem em uso, conforme Costa Val [4].

realizar análise discursiva manipula informações advindas das relações entre as *unidades*, como demonstra a autora, mas não se restringe a elas. É necessário, explicar as relações subjacentes no âmbito destas *unidades*, isto é, as relações discursivas. Desta forma, diferença entre texto e discurso faz-se relevante em nossa proposta, visto que a metodologia em questão propõe a articulação dos níveis textuais e discursivos, conforme explicitamos na sequência deste capítulo.

Neste estudo, *Texto* é identificado como a realização superficial de um discurso, reconhecendo, dessa forma, *Discurso* como uma entidade complexa composta por proposições relacionáveis pelas quais o autor/produtor/escrtior[1] expressa seu ponto de vista, seus objetivos. A forma como um autor materializa esse objetivo ocorre por meio de uma estrutura textual, na qual encontram-se reconhecidos os mecanismos e os elementos linguísticos que, a partir das suas relações, colaboram para composição da forma e do sentido da estrutura discursiva.

Relacionado à definição apresentada acima, corroboramos à definição proposta por Pardo [38], a qual expressa que:

um texto possui uma estrutura subjacente altamente elaborada que relaciona todo o seu conteúdo, atribuindo-lhe coerência. A essa estrutura dá-se o nome de estrutura discursiva, sendo ela objeto de estudo da área de pesquisa conhecida como Análise de Discurso.

[1] Autor /Produtor/Escritor: os termos são usados como sinônimos neste trabalho.

Ao construir um determinado texto, um autor utiliza-se de estratégias linguísticas específicas de acordo com o seu objetivo e em função do público a quem se destina a sua produção. Tais estratégias podem ser identificadas e estudadas no nível da microestrutura, em relações localizadas entre as estruturas na superfície do texto (intra e intersentencial); e no nível da macroetrutura, isto é, na organização semântico-conceitual inerente ao discurso. A compreensão de como se processa a relação entre os níveis micro e macroestrutural possibilita-nos reconhecer uma determinada estrutura discursiva como válida em termos formais e conceituais, no tocante à coesão e à coerência.

Conforme apresentamos, a investigação que propomos tem como fonte de estudo a estrutura textual, em que analisamos as relações microestruturais que se manifestam entre os segmentos localizados nos parágrafos e entre eles. Todavia, a análise realizada não se restringe unicamente às relações locais de superfície e a sua significação localizada. O objetivo desta análise é avaliar como os segmentos relacionam-se, o papel desempenhado e o resultado dessas relações na composição do sentido do discurso, bem como, podem identificados automaticamente sem a interferência humana.

Outro ponto de vista considerado na metodologia diz respeito da relação *Texto – Discurso* apresentada pelos autores Koch e Travaglia [20]. Para os autores, um texto é coerente se é possível lhe atribuir um sentido global em uma situação de comunicação. Em nossa proposta, consideramos que para haver um sentido global, deve ser possível determinar relações entre os conteúdos expressos pelos segmentos textuais, ou seja, entre suas proposições, em um sentido abstrato, manifestadas ao longo da superfície textual. É no nível do discurso que um escritor, ao produzir um texto, organiza e relaciona as proposições a fim de atingir determinados objetivos comunicativos, isto é, satisfazer suas intenções, como persuadir o leitor a realizar uma ação ou informar o leitor sobre algo.

Conforme observado, ao longo desta seção, há diversos aspectos de ordem composicional, seja ela formal ou conceitual, que se encontram envolvidos na produção, recepção e interpretação de um texto em uma determinada situação sociocomunicativa; aspectos estes que, em geral, não são percebidos isoladamente pelo receptor durante a sua articulação ou leitura. Considerando as observações feitas a respeito das teorias que envolvem *texto/discurso*, buscamos, construir uma metodologia específica passível de ser sistematizada, isto é, que pudesse constituir uma base heurística capaz de:

- identificar, categorizar e organizar as estruturas que constituem as partes de um texto, bem como a sua relação com a sua completude, relativizando os objetivos comunicativos e as intenções do autor ao produzi-lo;

- identificar as relações existentes entre as proposições, sejam elas da ordem informativa/intencional/argumental;

- identificar o objetivo do texto em uma estrutura condensada, capaz de representar sinteticamente o que está manifestado através dos constituintes textuais,

Na perspectiva metodológica elaborada no âmbito desta investigação, a estrutura textual é analisada como um todo relacionado, um conjunto de unidades formais e conceituais hierarquicamente relacionadas, coesas e coerentes. Neste sentido, para analisarmos esse *todo significativo* contamos com diferentes níveis de informações, isto é, consideramos características sintáticas, marcadores discursivos, sinais de pontuação, bem como, informações semânticas, as quais são identificadas a partir das relações retóricas estabelecidas entre os segmentos ao longo de toda a estrutura.

Desta forma, trazendo à luz a proposição de Vygotskij [44], ao explicar a relação entre pensamento e linguagem; observamos linguagem no sentido de ocorrência textual, em que *o significado de uma palavra é uma potência que se realiza no discurso vivo na forma de sentido*, parafraseamos a afirmação do autor, ao associarmos que a palavra está para o texto, como pensamento está para discurso. É neste sentido que as definições de texto e discurso sobrepõem-se no processo de análise previsto nesta tese.

3.2.2 Elementos da Relação Texto – Discurso

A proposta metodológica, desenvolvida e apresentada nesta tese, tem como objetivo realizar a análise automática de um discurso. O texto é o objeto de investigação, a partir do qual toda a sua estrutura é avaliada em termos formais até identificação da sua orientação discursiva. Na sequência, reproduzimos um texto que faz parte de um dos *corpora* que constituem essa investigação. Especificamente, o texto apresentado encontra-se no corpus do Jornal *Público* do ano 1994 e faz parte do conjunto aprendizado, a partir do qual foram propostas regras de análise textual, conforme a figura 3.2.

Alguns elementos e definições, tais como *proposições, segmentos e subsegmentos, núcleo e satélite e, macroprosição* pertinentes à relação *Texto – Discurso*, são apresentados e esclarecidos na continuidade deste capítulo, para a compreensão da proposta.

Morreu «Sonny» Constanzo.

Dominic «Sonny» Constanzo, que acompanhou, com o seu trombone, cantores como Ella Fitzgerald e Tony Bennett, morreu quinta-feira, em New Haven, no estado americano do Connecticut, aos 61 anos, depois de um transplante cardíaco. Ao longo da sua carreira tocou com o clarinetista Woody Herman, o trompetista Thad Jones, o baterista Mel Lewis e o cantor Clark Terry. Durante muito tempo acompanhou a vocalista Rosemary Clooney, à frente da sua grande orquestra. Mas apenas em 1992 conseguiu fazer a primeira gravação para uma grande etiqueta, no caso a Stash.

Figura 3.2: Exemplo de um texto selecionado no Jornal Público 1994 – publico-19940101-007.

3.2.2.1 As Proposições

Conforme apresentamos, o texto é a unidade de estudo nesta investigação enquanto estrutura formal, constituído por um conjunto de segmentos e subsegmentos interrelacionados, os quais representam concretamente as proposições, formando toda estrutura discursiva.

No que se refere à definição para o termo *Proposição*, apesar de uma extensa busca na literatura, não existe uma definição precisa e única para a expressão, quando se pretende aplicá-la a estudos interdisciplinares, que envolvem áreas completares como é o caso da presente pesquisa.

A definição apresentada em dicionários para o termo provém do Latim e designa o ato ou o efeito de propor algo; realizar uma proposta, um oferecimento; uma estrutura que é apresentada por alguém para se chegar a uma conclusão. Além disso, a idéia associada ao termo *proposição* pode também ser identificada como a expressão verbal de um juízo; uma oração gramatical; uma parte do poema em que se indica o assunto de que se vai tratar; um teorema.

Pardo [38] realiza uma minuciosa investigação a respeito da definição do termo. Segundo o autor, de acordo com a literatura especializada, os termos *cláusula*, *segmento textual*, *segmento discursivo*, *trecho de texto* e *proposição*, entre outros, possuem diferentes significados. Entretanto, no contexto das pesquisas de análise automática de discurso, esses termos têm sido usados de forma indiscriminada e o próprio autor, justificando-se, emprega os termos, em seu trabalho, de forma intercambiável.

A definição que utilizamos, nesta tese, para determinar o que entendemos e reconhecemos por *proposição* é mesma definição utilizada pela linha de investigação da análise do discurso. Para os analistas do discurso, as proposições são estruturas ou entidades abstratas, uma abstração que representa um conceito, uma idéia que se realiza através de segmentos. Salientamos que, quando se trabalha no nível do discurso, como é o caso da metodologia proposta, o termo *proposição* ou *segmentos discursivos* encontra-se relacionado ao conteúdo representado por um segmento textual de qualquer extensão.

No caso da presente investigação, as proposições, representadas pelos segmentos, que constituem os textos em análise, são determinadas automaticamente a partir do resultado da aplicação de um conjunto de regras, desenvolvidas exclusivamente para compor essa etapa de análise, considerando-se as informações e características apresentadas pela análise do *Palavras* nos textos dos *corpora*, para a sua constituição, conforme descrevemos no capítulo 4. No exemplo apresentado na figura 3.3, é possível evidenciar o início de uma proposição a partir da sigla *UTT* - (utterance), enquanto que o seu limite é dado pelo ponto final ou pelo início de um novo *UTT* - (utterance), demarcados automaticamente pelo analisador *Palavras*.

No exemplo apresentado na figura 3.3, apresentam-se delimitadas as duas primeiras proposições do texto, considerando-se para essa demarcação as informações sintáticas geradas pelo próprio analisador automático *Palavras*. Optou-se por determinar as proposições dos textos a partir do resultado do analisador sintático, a fim de garantir uma padronização na identificação das proposições. Além disso, o resultado do analisador *Palavras* oferece características que associadas a outras informações formais/conceituais possibilitam automatizar uma etapa do processamento discursivo que, até então, é realizada por poucos analisadores discursivos. As duas primeiras proposições do texto analisado são apresentadas na figura 3.4.

No exemplo apresentado na figura 3.5 é possível identificar as demais proposições referentes ao texto analisado pelo *Palavras*. O resultado da análise sintática automática apresenta o texto esquematizado a partir do que definimos em nossa investigação como *Proposições*. Assim, o texto analisado e exemplificado é composto por cinco proposições de diferenciadas extensões, sendo essas representadas concretamente por um ou vários segmentos e subsegmentos.

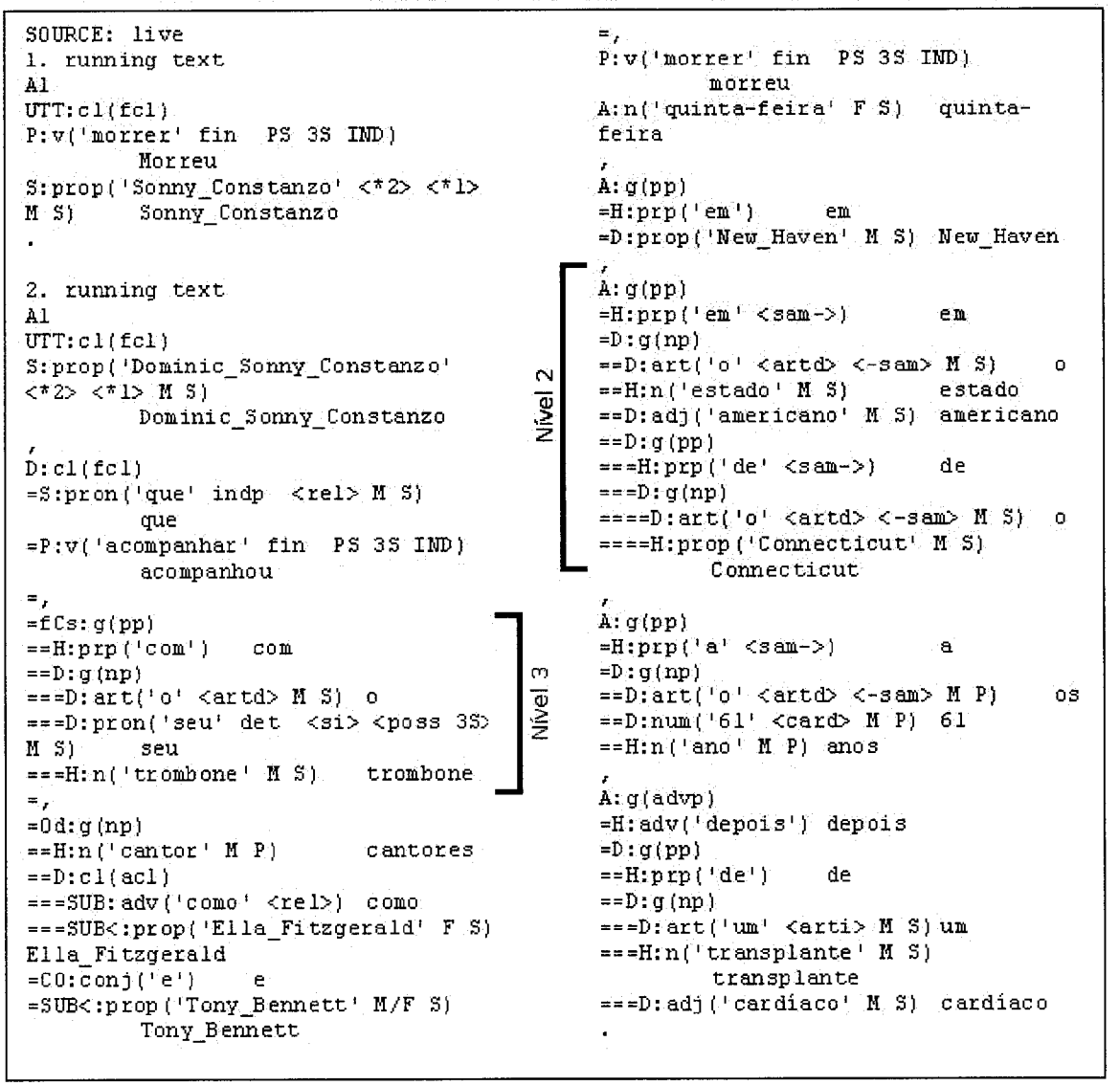

Figura 3.3: Exemplo referente à análise automática (parcial) realizada pelo *Palavras*, texto Jornal Público 1994.

No tocante ao processo de automatização relativo à identificação das proposições do discurso, ressaltamos que esse processo não é considerado como um módulo autônomo proposto em nossa metodologia, mas é parte constituinte da primeira etapa do processamento do *AuTema-Dis*. Em referência à identificação e à delimitação automática das proposições do discurso, a partir da execução do *Palavras*, avaliamos satisfatoriamente os resultados obtidos, visto que, permitiu-nos a utilização desses resultados na constituição da base do processamento, excluíndo-se as intervenções dos analistas que, normalmente, são realizadas de maneira arbitrária e subjetiva.

Proposição 1
Morreu «Sonny» Constanzo. [título]

Proposição 2
Dominic «Sonny» Constanzo, que acompanhou, com o seu trombone, cantores como Ella Fitzgerald e Tony Bennett, morreu quinta-feira, em New Haven, no estado americano do Connecticut, aos 61 anos, depois de um transplante cardíaco.

Figura 3.4: Exemplo das duas primeiras proposições apresentadas a partir da análise automática realizada pelo *Palavras* em um dos textos do corpus do jornal Público 1994.

Proposição 3
Ao longo da sua carreira tocou com o clarinetista Woody Herman, o trompetista Thad Jones, o baterista Mel Lewis e o cantor Clark Terry.

Proposição 4
Durante muito tempo acompanhou a vocalista Rosemary Clooney, à frente da sua grande orquestra.

Proposição 5
Mas apenas em 1992 conseguiu fazer a primeira gravação para uma grande etiqueta, no caso a Stash.

Figura 3.5: Exemplo apresenta as demais proposições identificadas a partir da análise automática realizada pelo *Palavras* no texto jornal Público 1994.

3.2.2.2 Os Segmentos e Os Subsegmentos

A definição da palavra *Segmento* que se encontra dicionarizada origina-se do Latim *segmentu* e é empregada para designar seção; uma parte de um todo; uma porção determinada de um objeto; um fragmento. As pesquisas em análise do discurso os definem como estruturas textuais que representam concretamente as proposições discursivas.

Em nosso trabalho, assumimos que os segmentos são as menores unidades de significação entre os quais são estabelecidas relações de significado que compõem estrutura discursiva. No desenvolvimento da pesquisa, observou-se que alguns segmentos desempenhavam *papéis* diferenciados na composição da estrutura textual, portanto, não apresentavam a mesma relevância em relação ao tema do discurso. Na metodologia que propomos para a automatização da análise discursiva e produção da macroestrutura/macroproposição de um texto, a distinção entre *segmento e subsegmento* é determinante para o resultado final da análise.

Os segmentos ocupam o *status* de primeiro nível na representação de uma proposição, pois encontram-se diretamente relacionados ao tema do texto. Os subsegmentos, por sua vez, estão indexados aos segmentos, constituintes de 1° nível, o que promove a existência relações hierárquicas de subordinação e, algumas vezes, de paralelismo, como é o caso dos processos de coordenação, conforme apresentamos no capítulo 2. Em relação aos subsegmentos foi necessário desenvolver uma definição no âmbito desta tese, para contemplar as exigências metodológicas.

Na perspectiva para análise discursiva que apresentamos, *Subsegmento* é definido como um tipo de segmento que desempenha um papel com um menor grau de relevância ou não contribui efetivamente à tematicidade da estrutura na qual fazem parte. Assim, identificamos subsegmentos como constituintes textuais que desempenham uma função complementar e acessória em relação aos *Segmentos*, existindo entre ambos uma relação semântico-conceitual, que pode ser da ordem informativa, argumental ou intencional, as quais contribuem para a tessitura do texto. Apesar de verificarmos que diferentes relações se estabelecem entre os segmentos e subsegmentos e auxiliam à composição temática, a investigação realizada, demonstra que alguns dos subsegmentos podem ser omitidos ou excluídos, em função do nível de profundidade em que se encontram, conforme apresentamos no capítulo 4.

Outrossim, dependendo do tipo de relação que se estabelece entre os constituintes, ou seja, relações do tipo informativa, argumental ou intencional, é possível determinar se um subsegmento pode ser mais ou menos relavante, passível de ser eliminado, por exemplo, no momento da geração automática da macroestrutura/macroproposição de um texto. Nota-se que tanto os segmentos quanto os subsegmentos são consideradas entidades concretas, analisadas na superfície do texto apesar de *representarem* o conceitual, a forma lógica.

Independente da classificação, *segmento ou subsegmento*, é de conhecimento acadêmico que a delimitação dessas unidades de significado representa uma séria dificuldade para os trabalhos que envolvem a segmentação textual. A dificuldade não diz respeito apenas ao reconhecimento dessas unidades, mas também em relação aos seus limites, as suas fronteiras

no escopo do texto.

Pardo e Nunes [39], a partir de suas pesquisas, confirmam a dificuldade em definir segmentos de forma precisa, Assim, para identificação dos segmentos em seus trabalhos, recorrem às regras produzidas por Marcu [25; 26], em que o autor propõe um conjunto de regras baseadas na ocorrência de sinais de pontuação do texto e marcadores discursivos. Na visão de Marcu, os sinais de pontuação e os marcadores são os principais indicativos superficiais da estruturação textual.

As regras sintáticas e os sinais de pontuação, conforme propõe Marcu [25; 26] são fiáveis na identificação dos segmentos, devidamente comprovado nos trabalhos referidos no capítulo 2 desta tese. Todavia, optamos, na composição da nossa metodologia, por recorrer a uma estratégia que, além de possibilitar a restrição na identificação dos segmentos, pudesse responder de forma automática a essa identificação e delimitação. Para tal, optamos por considerar informações advindas de um analisador automático, o *Palavras*, conforme apresentamos no capítulo 4 desta tese, identificadas em um conjunto de regras.

Pardo [38] desenvolveu um primeiro analisador discursivo para o Português Brasileiro, o *DiZer*. É importante ressaltar que para realizar a identificação dos segmentos e dos seus limites, o *DiZer* considera os critérios de pontuação e específicas marcas na superfície do texto. Na análise realizada pelo DiZer, o início de um segmento coincide com o início de uma sentença e o seu limite final pode ser determinado ao encontrar uma vírgula, ponto e vírgula e dois pontos; ou marcadores textuais fortes.

No caso da nossa investigação, a pontuação, em situações específicas, também é utilizada para a identificação e delimitação das fronteiras dos segmentos e subsegmentos, mas não de forma definitiva. O reconhecimento dos constituintes é feito automaticamente, considerando-se os dados e as características resultantes do processo realizado pelo analisador *Palavras*, os quais são passíveis de serem convertidas em regras restritivas, que devidamente sistematizadas realizam automaticamente a identificação, a delimitação e a determinação das fronteiras dos segmentos e subsegmentos, conforme apresentamos a seguir no capítulo 4.

A figura 3.6 é representativa da totalidade dos segmentos e subsegmentos identificados na 3ª proposição de um dos textos dos *corpora*, utilizado nesta investigação. Os constituintes foram determinados partir da aplicação de regras de segmentação propostas com algumas características advindas do resultado do analisador *Palavras* e outras de ordem estrutural e semântica, quando possível.

Semelhante à figura 3.6, a figura 3.7 apresenta 2ª proposição do texto em que aparecem os

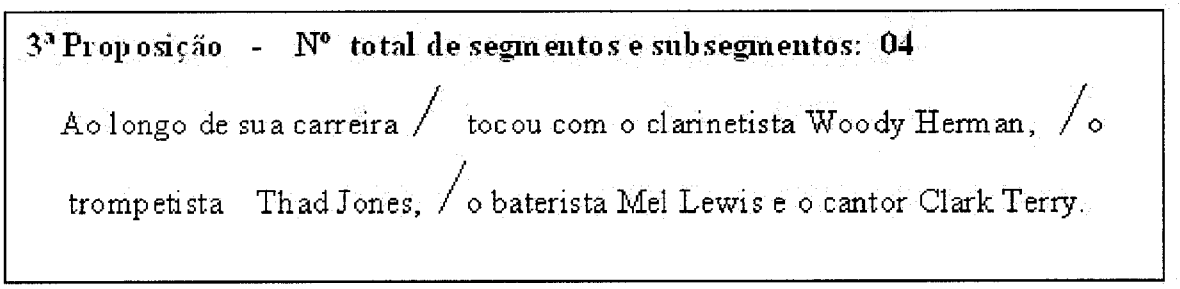

Figura 3.6: Exemplo dos segmentos e subsegmentos identificados automaticamente na 3ª proposição do texto – publico-19940101-007.

segmentos e os subsegmentos delimitados. No entanto, neste estágio de análise do texto, os constituintes não recebem a classificação, isto é, o analisador *AuTema-Dis* não apresenta a distinção entre segmento e subsegmento e nem os níveis em que eles se encontram na estrutura, trata-se da etapa inicial de reconhecimento e identificação da proposição e dos seus elementos a partir do processamento do *Palavras*. A etapa posterior, a ser processada no módulo seguinte, realiza a classificação d os constituintes, conforme apresentamos na sequência desta tese, especificamente no capítulo 4.

Nas imagens apresentadas nas figura 3.6 e figura 3.7 com a identificação dos segmentos e subsegmentos é possível que o leitor seja levado a acreditar que a apresentação dos constituintes está condicionada apenas por critérios de pontuação, o que não é correto. Todavia, ressaltamos que as figuras representam a etapa inicial de reconhecimento dos elementos textuais, não havendo filtragem e nem classificação dos constituintes, o que vem a ser realizado na etapa seguinte, constituída pelas regras de segmentação e delimitação, conforme apresentamos na metodologia.

3.2.2.3 Segmentos–Subsegmentos e a Concepção de Núcleo–Satélite

Os segmentos que compõem a estrutura textual desempenham papéis diferenciados na composição da estrutura textual. Inicialmente, poder-se-ia relacionar a classificação Segmento – Subsegmento à concepção de *Núcleo – Satélite* apresentada por Mann e Thompson [23], no entanto, a conceituação não se sobrepõe, procuramos apresentar os pontos em que as definições se aproximam e se relacionam, considerando o tipo de abordagem que propomos.

Conforme apresentamos na subsubseção 3.2.2, os Segmentos e Subsegmentos são responsáveis pelas relações de coesão e pela coerência, no tocante à apresentação e evolução do tema ao longo do texto. Em relação à metodologia proposta e do tipo de análise automática pretendida, optou-se por classificá-los em função da sua interação e comprometimento com a configuração do tema do discurso, o que se afasta, em termos estruturais, da proposta da RST.

Proposição 2

1° Segmento – Dominic «Sonny» Constanzo,

2° Segmento – que acompanhou,

3° Segmento – com o seu trombone,

4° Segmento – cantores como Ella Fitzgerald e Tony Bennett,

5° Segmento – morreu quinta-feira,

6° Segmento – em New Haven,

7° Segmento – no estado americano do Connecticut,

8° Segmento – aos 61 anos,

9° Segmento – depois de um transplante cardíaco.

Figura 3.7: O exemplo apresentado é representativo dos segmentos e subsegmentos referentes a 2ª proposição do texto, identificados pela análise automática do *Palavras* em um texto jornal Público 1994.

No capítulo 2, apresentamos a proposta RST, em que os autores determinam que o Núcleo - apresenta a informação mais importante e o Satélite - apresenta uma informação que complementa o que é apresentado no Núcleo, existindo uma relação retórica entre eles. Neste sentido, a nossa proposta se aproxima à dos autores, isto é, o *subsegmento*, desempenha um papel complementar, subordinado em relação ao *segmento*, podendo também existir uma relação retórica entre eles.

Para os autores, não existe uma ordem preestabelecida para a manifestação dos elementos *N* e *S* no texto, mas existem restrições na maneira com os Núcleos e os Satélites se relacionam. Nas restrições apresentadas, os autores determinam que um Núcleo pode se relacionar com um ou vários Núcleos ou com um Satélite. No entanto, os Satélites não apresentam essa característica, isto é, eles só podem estabelecer relações com Núcleos. As restrições apresentadas servem para classificar as relações em dois tipos: *nuclear*- quando se trata de uma relação N-S e *multinuclear*- quando existe a relação entre núcleos.

No caso da metodologia que propomos, as relações retóricas entre *segmentos* não são contempladas, pois não tínhamos como objetivo identificar este tipo de relação, assim, justificamos não ter desenvolvido regras para explicar as relações entre *segmentos*. Um outro fator fator para não atribuição de relações entre os *segmentos* está relacionado às restrições estabelecidas pelas regras propostas para segmentação textual, todavia, as relações internas aos subsegmentos são identificadas em nossa proposta, conforme apresentamos no capítulo 5, na figura 5.2.

A possibilidade de *relação* entre o *N e S* definida na *RST* é revisada em nossa proposta. Nesta investigação, os critérios de *relacionamento* entre os segmentos e subsegmentos são definidos em função da sua representatividade em relação ao tema. Acreditamos que existe a possibilidade de ocorrer uma relação retórica entre *subsegmentos*, o que é proibitivo pela *RST*. Além disso, dependendo do nível de profundidade em que se encontra um subsegmento, talvez não seja necessário e nem relevante atribuir uma relação retórica entre os constituintes, podendo o subsegmento ser descartado.

Consideramos, em nossa arquitetura, a alternativa de atribuirmos relações retóricas entre *subsegmentos* condicionada à relevância que estes constituintes detém em relação ao tema do texto. Além disso, acreditamos, pelos resultados da investigação realizada, que existe entre os *subsegmentos* um relação hierárquica, em que um subsegmento que ocupa, por exemplo, o 3° nível em uma árvore DTS seja subordinante em relação a outro *subsegmento* que ocupa o 4° nível, o que justificaria a possibilidade de atribuirmos de uma relação retórica entre eles ou excluí-lo, no momento da elaboração da macroestrutura/macroproposição. Todavia, a diferença entre as propostas recai na possibilidade de exclusão de alguns subsegmentos e atribuição, os quais não se encontram diretamente relacionados ao tema, o que está em conformidade com o que descreve Van Dikj [10], quando propõe as *macrorregras* de supressão e de generalização para a construção da macroestrutura.

Um ponto de aproximação importante nas relações *segmento – subsegmento* e *núcleo – satélite* é noção de hierarquia textual entre os constituintes, pois está diretamente relacionada à identificação da *macroproposição* e a sua representação através de uma *macroestrutura*. A hierarquia observada na organização textual fornece dados para a construção de árvores de dependência dos segmentos *DTS's*[2], através das quais identificamos os segmentos e subsegmentos hierarquicamente dispostos. Essa identificação arbórea é importante para determinar quais são os segmentos principais e quais são os segmentos acessórios ou secundários, isto é, os subsegmentos, em relação à cadeia temática, bem como, quais dos subsegmentos que

[2]DTS: Árvore de Dependência dos Segmentos: o conceito foi desenvolvido no âmbito desta tese e serve para designar a forma de representar e apresentar os segmentos e dos subsegmentos de um texto em um esquema arbóreo. Nas DTS's são considerados os princípios que orientam hierarquização da informação na estrutura textual.

não contribuem efetivamente para a constituição do tema.

3.3 Macroestrutura–Macroproposição – uma relação global de significação –

Nossa proposta metodológica é constituída por quatro módulos autônomos que realizam análise textual a partir das relações que se estabelecem entre os segmentos que compõem toda a estrutura, relações microestruturais. Pautada na perspectiva de avaliar completamente um texto, a metodologia em questão tem como objetivo final da análise apresentar uma estrutura sintética representativa da macroestrutura do texto avaliado, isto é, a *Macroproposição*.

Segundo Favero e Koch [13], um texto consiste em qualquer passagem, falada ou escrita, que forma um todo significativo, independente de sua extensão. Trata-se pois, de uma unidade de sentido, de um contínuo comunicativo contextual que se caracteriza por um conjunto de relações responsáveis pela tessitura do texto. Da afirmação apresentada pelos autores, a parte relevante ao nosso trabalho está relacionada à identificação da *tessitura*, pois a produção de textos coerentes pressupõe esse *tecido textual*, considerando *a priori* a articulação entre os níveis micro e microestruturais, que se manifestam a partir das relações entre os seus segmentos na composição do todo significativo.

A metodologia que propomos executa uma análise que contempla inicialmente as relações no nível microestrutural, identificado como o nível da coesão. É neste nível que se encadeiam relações de diferentes tipos entre seus segmentos, que constituem a superfície textual. É, também, neste nível que as relações entre as estruturas favorecem e propiciam a distribuição hierárquica das informações ao longo de toda a estrutura, contribuindo, desta forma, na apresentação coesa e coerente do tema. As relações observadas neste nível recebem o nome de relações retóricas, e são avaliadas, nesta investigação, sob a perspectiva apresentada pela *RST* [23], conforme apresentamos nos capítulos 2 e 4 desta tese.

Em relação à compreensão global de um texto, considerando-o a partir de uma base linear, é necessário perceber que, duas proposições estão conectadas entre si se é possível denotar-se uma relação de similaridade referencial entre elas. Na relação entre proposições, os fatos que elas denotam encontram-se ligados, desde que um seja condição de outro. Portanto, as relações microestruturais entre os segmentos são da ordem formal e conceitual. Desta forma, o texto é uma unidade significativa de coerência que requer o equilíbrio entre a continuidade temática e a progressão semântica, integrando as sequências microestruturais constituindo,

de forma coesa e coerente as relações macroestruturais.

A etapa inicial da metodologia realiza a análise das relações microestruturais. É no nível microestrutural que as sequências frásicas encadeiam respectivas *ações*. Para Van Dijk [7], as frases topicais são estruturas cujo conteúdo proposicional corresponde aproximadamente ao conteúdo da macroestrutura da sequência, adequando-se, assim, ao que está previsto na metodologia que propusemos. Neste sentido, a estrutura esquemática textual controlaria a formação das macroproposições locais e determinaria, por exemplo, se em um processo de sumarização, o texto está completo ou interrompido e que tipo de informação cada segmento requer. As macroproposições locais ordenadas estabelecem a formação das sequências de frases do texto.

Em um dos seus primeiros trabalhos, Van Dijk [5], avaliou o processo de criação de resumos de histórias, a partir dos resultados propõe que a construção da macroestrutura textual seja um elemento essencial para a compreensão de qualquer texto. A macroestrutura é uma estrutura de significação global de um texto e seria derivada da microestrutura ou base de texto. A macroestrutura pode, desta forma, ser identificada como a estrutura semântica subjacente ao texto, fruto de um processo de *sumarização*, realizado através da aplicação de regras de redução de informação semântica.

Assim, no âmbito desta tese, identificamos *macroestrutura* a partir da definição apresentada por Van Dijk [11], caracterizada como uma espécie de estrutura profunda semântica do texto; que dá conta do conteúdo do mesmo. Associada à noção de macroestrutura está a sua representação, isto é, o conceito de *macroproposição*; estrutura que é obtida através da aplicação de *macrorregras*. O exemplo que segue apresenta o texto na sua construção original e sua macroprosição, conforme a figura 3.8.

As macrorregras reduzem e abstraem o conteúdo proposicional das sequências textuais, além de organizarem o seu conteúdo em termos hierárquicos. As *macrorregras* foram apresentadas por Van Dijk [10], identificadas como: (a) supressão ou apagamento; (b) generalização; (c) construção. Tais regras chamam-se macrorregras porque produzem macroestruturas. A sua função consiste em transformar a informação semântica e a frase que expressa em uma macroestrutura, representada por uma macroproposição. Especificamente em nossa investigação, conforme apresentamos no capítulo 4, a metodologia proposta faz uso de apenas uma das macrorregras, isto é, supressão ou apagamento.

Van Dijk [10] apresenta essas macrorregras como operações que selecionam, reduzem, generalizam e (re)constróem proposições em outras proposições menores, mais gerais ou mais particulares. Elas, segundo o autor, são regras de interpretação de sentenças e de pares

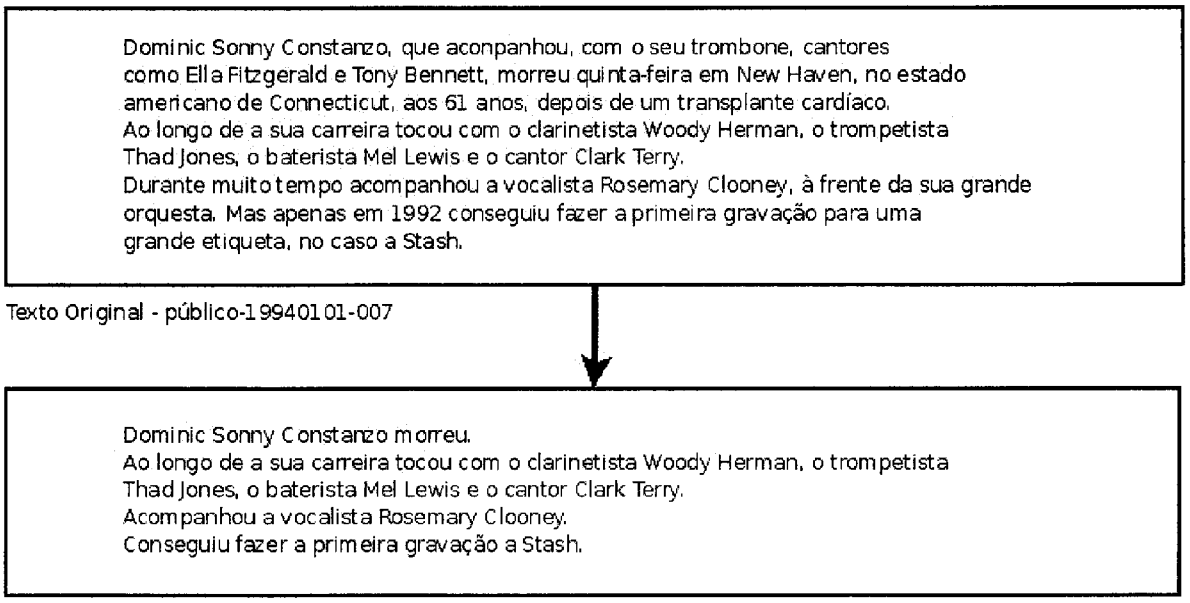

Figura 3.8: A figura representa o texto completo e sua respectiva macroestrutura/macroposição - publico19940101-007.

de sentenças como proposições (globais), que caracterizam o significado de uma sequência de ações realizadas. Assim as macrorregras suprimem toda a informação proposicional de relevância exclusivamente local que não seja necessária para a compreensão do resto do discurso. Certamente, essas macrorregras podem operar somente com base no conhecimento do mundo, ou em um conceito de mundo restrito, (re)criado, por exemplo, o que pode ser identificado em ontologias [3].

As macroestruturas derivam das microestruturas, mas não é sempre que podemos efetuar a operação inversa, pois, para isso, precisamos saber diferenciar as macrorregras de apagamento e de generalização das macrorregras de construção e de integração. As macrorregras podem ser aplicadas muitas vezes, mas essa recursividade passa a ser limitada pelo princípio da informatividade. Sua aplicação está condicionada aos esquemas determinados pelo produtor do texto, que nos permitem estabelecer inferências necessárias para a construção da macroestrutura textual. Ao propormos a metodologia para análise automática do discurso, foi previsto a execução em sistema, neste sentido, optou-se por fazer uso da macrorregra de

[3]Ontologia: o termo é identificado, nesta investigação, conforme o sentido atribuído computacionalmente. Em Ciência da Computação e Ciência da Informação, uma ontologia é um modelo de dados que representa um conjunto de conceitos dentro de um domínio e os relacionamentos entre estes. Uma ontologia é utilizada para realizar inferência sobre os objetos do domínio. É uma forma de representação de conhecimento sobre o mundo ou alguma parte deste. Ontologias geralmente descrevem: * Indivíduos: os objetos básicos; * Classes: conjuntos, coleções ou tipos de objetos; * Atributos: propriedades, características ou parâmetros que os objetos podem ter e compartilhar; * Relacionamentos: as formas como os objetos podem se relacionar com outros objetos.

apagamento ou supressão, em função das limitações de implementação e execução computacional da análise.

Conforme evidenciamos em nossa investigação, a macroestrutura é definida ao nível da representação semântica global do texto, apresentando como correlato psicológico, um esquema cognitivo que determina a planificação, execução, compreensão e reprodução de um texto. A utilização de estratégias linguísticas, tais como, as macrorregras, permitenos adequar uma determinada construção a uma superestrutura [4] particular, adequado a sua função sociocomunicativa e semântica.

Cabe-nos ressaltar um ponto relacionado à identificação da macroestrutura e à produção da sua macroproposição, isto é, a concepção de *Superestrutura.* Nesta tese, concebemos as superestruturas como formas ou esquemas globais de categorias funcionais culturalmente convencionadas, que respeitam determinadas regras de combinação e ocorrência, as quais impõem certas restrições semânticas.

Van Dijk [9] esclarece que os tipos de textos se diferenciam não apenas por suas diferentes funções comunicativas, por seus diferentes tipos de conteúdos e por suas diferentes funções sociais, mas também possuem diferentes tipos de construção. Dessa forma, os textos não apenas possuem uma estrutura semântica global, possuem também uma estrutura esquemática global, chamada de *superestrutura.* Uma estrutura esquemática consiste em uma série de categorias hierarquicamente ordenadas, muito similares às categorias de um esquema narrativo. Estas categorias possuem funções específicas relacionadas às respectivas macroproposições de um texto. Uma superestrutura esquemática é meramente uma estrutura formal, muito próxima à sintaxe de uma oração. Ela é preenchida com o conteúdo da macroestrutura semântica. Por exemplo, em princípio, qualquer discurso jornalístico apresenta o mesmo esquema de notícias, mas o conteúdo global do texto é diferente em cada caso, o que foi evidenciado neste estudo.

Texto, nesse sentido, pode ser reconhecido como uma grande proposição, a partir da qual se pode identificar uma macroestrutura, a ser representada por uma - *macroproposição* de sentido completo. Em uma macroproposição observam-se grupos de proposições de menor tamanho, que se relacionam nos níveis morfológico, sintático, semântico e pragmático, a fim de compor uma estrutura representativa do discurso. A fim de exemplificação esquemática, apresentamos a figura 3.9 representativa da interação dos níveis que compõem uma estrutura textual.

[4] Superestrutura: o conceito de superestrutura foi desenvolvido por Dijk [6]. Para o autor, as superestruturas textuais são estruturas globais que se assemelham a um esquema. Elas delimitam a forma global do texto em termos de categorização esquemática.

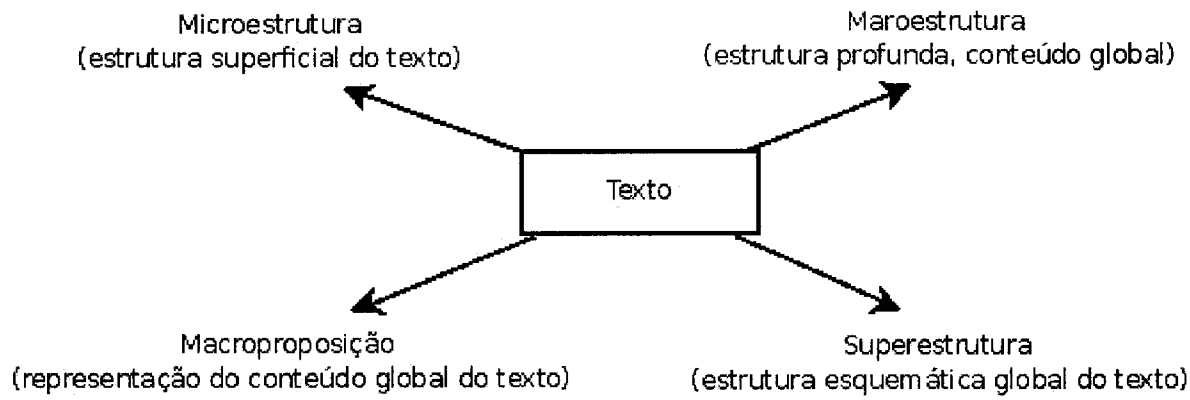

Figura 3.9: Relações entre os níveis textuais.

A metodologia proposta estabelece a articulação e relação de dois níveis, o micro e o macro-estrutural e, considerando tal articulação é possível, entre outras coisas, produzir uma estrutura organizada que representa semanticamente toda a macroestrutura do texto analisado, referimo-nos à *macroproposição*.

3.4 Elementos complementares à análise automática

A proposta metodológica para análise automática do discurso é composta, na sua base, por informações da ordem morfo-sintático-semântica advindas do resultado da análise realizada, em uma primeira etapa de avaliação, pelo analisador *Palavras* nos textos dos *corpora*. O processo de identificação das proposições nos textos, bem como, o processo que reconhece e classifica os segmentos e os subsegmentos e a própria atividade de segmentação são produzidos com essas informações e características advindas da análise sintática.

Todavia, em casos específicos, em que as informações e características da análise sintática do *Palavras* tornam-se insuficientes para dar conta da identificação dos segmentos e realizar a segmentação do texto, a metodologia também prevê que sejam utilizadas informações para auxiliar nos processos mencionados. Existem estudos realizados em *PLN*, especificamente em sumarização automática e análise automática de discurso, tais como, os de Marcu [2] e Pardo [38], que recorrem aos marcadores discursivos, aos sinais de pontuação e algumas regras sintáticas específicas para segmentação, associadas à possibilidade de identificação das relações retóricas, etc.

No caso da metodologia em questão, identificamos os *elementos complementares* da seguinte forma:

3.4.1 Sinais de Pontuação

No caso da nossa metodologia, os sinais de pontuação são considerados a partir da identificação realizada pelo analisador *Palavras*, conforme descrevemos no capítulo 4 desta tese. Os sinais de pontuação são considerados como indicativos de segmentação e estão incluídos no conjunto das regras que identificam e realizam a segmentação textual. Salientamos, todavia, que os indicativos de pontuação não são os elementos que determinam os segmentos e os subsegmentos, não são eles os responsáveis pelo processo de segmentação dos constituintes textuais. No entanto, a pontuação, em algumas circunstâncias, em que, por exemplo, o analisador *Palavras* não consegue demarcar os constituintes, recorre-se a uma das regras que contém a caracterização da pontuação, para auxiliá-las no processo de segmentação das unidade textuais.

Em casos específicos, os sinais de pontuação constituem uma regra para *não segmentação*, isto é, dependendo das estruturas, tais como, interrogativas, exclamativas e as citações/afirmações iniciadas a partir dos dois pontos até o ponto final, a metodologia proposta determina que essas construções devem constituir um único segmento, não havendo possibilidade de serem identificados diferentes subsegmentos ou ocorrer algum tipo de subsegmentação no seu interior.

Além dos sinais de pontuação, incluimos na mesma regra de *não segmentação* os sinais de intercalação textual: (); ; [], bem como, estruturas que se encontram entre travessões.

3.4.2 Marcadores Discursivos

Os marcadores discursivos apoiam a nossa metodologia no processo de identificação dos segmentos, classificação e segmentação das estruturas, semelhante à utilização que é feita considerando-se os sinais de pontuação. Na metodologia desenvolvida não há regras específicas que contemplem o processo de segmentação a partir das características exclusivas dos marcadores discursivos, como se pode, por exemplo, observar em Pardo [38], trabalho em que o autor afirma que os marcadores discursivos são os maiores indicadores de relações retóricas de uma estrutura discursiva de um texto, apoiando toda a segmentação textual na identificação destes elementos.

A nossa abordagem identifica e reconhece que os marcadores discursivos são indicadores da existência de uma relação retórica entre partes de um texto, mas não restringe a possibilidade de identificar uma relação, considerando apenas esses elementos. Os marcadores são reconhecidos, primeiramente, na análise realizada pelo *Palavras* e, na sequência da avaliação, são incluídos no conjunto das heurísticas propostas para identificar os segmentos, realizar a segmentação e atribuir, quando possível, uma relação retórica entre um segmento e um subsegmento.

Salientamos que na metodologia proposta não foi desenvolvida nenhuma regra específica para contemplar os casos dos marcadores, existem regras com características morfo-sintático-semânticas que reconhecem determinado marcador como uma das possibilidades de atribuir uma relação retórica, como podemos evidenciar com a conjunção de oposição/opositiva *Mas*.

3.4.3 Regras Sintáticas

Nossa metodologia está assente em questões de ordem morfo-sintática-semânticas. As regras que compõem um dos módulos de análise da metodologia proposta de análise discursiva foram edificadas com base nas informações da análise sintática realizada pelo *Palavras* em textos dos *corpora*, conforme apresentamos no capítulo 4.

As heurísticas que compõem a nossa metodologia estão destinadas à segmentação os textos e, quando possível, à atribuição de relações retóricas entre os seus segmentos e subsegmentos. Há um formalismo na organização das regras que permite a sua automatização na realização do processo.

Marcu [2] desenvolveu um conjunto de regras sintáticas para realizar a segmentação manual de textos escritos em Inglês. Algumas das regras propostas pelo autor foram também identificadas em nossa abordagem a partir das características evidenciadas nos *corpora* da nossa pesquisa. Um exemplo de equivalência entre as regras de Marcu e as que compõem a nossa metodologia é a regra que exclui a possibilidade das estruturas, que desempenham papel de sujeito ou de objeto de um determinado verbo, constituírem unidades isoladas do discurso ou, no caso da nossa abordagem, serem reconhecidas como um segmento independente.

Ressaltamos que as regras sintáticas constituem parte da metodologia que propomos para automatização da análise discursiva. Elas foram elaboradas a partir da análise dos *corpora* que constituem essa investigação, conforme apresentamos no capítulo referente à metodologia.

3.4.4 Categorização Verbal

Os elementos verbais têm um papel significativo no âmbito de qualquer estudo que avalie as relações estruturais e conceituais do discurso, neste sentido, a identificação deste constituinte se faz relevante em nosso estudo. A avaliação do papel do elemento verbal no processo de segmentação textual e na identificação da macroproposição é fundamental está inserida no conjunto das regras propostas tanto para a segmentação quanto para a apresentação/representação do tema do texto analisado.

As informações a respeito dos verbos presentes nos textos analisados provém da análise automática realizada inicialmente pelo *Palavras*, as informações e as características atribuídas pelo analisador são consideradas para a constituição das regras. Não há, em nossa proposta metodologia, regras específicas com base apenas na informação verbal, que estão destinadas a realizarem a segmentação ou a organização da macroestrutura/macroproposição. Procuramos, neste sentido, desenvolver um conjunto de regras para ambas atividades que pudessem utilizar as informações categoriais dos verbos para uma melhor performance no tocante à execução de um sistema computacional, conforme apresentamos no capítulo 4.

Ressaltamos que as informações categoriais e argumentais que são atribuídas aos verbos, as quais estão relacionadas à questão da complementação estrutural pode ser revista em um trabalho futuro, pois acreditamos que existem muitas possibilidades a serem desenvolvidas neste sentido, que poderiam ajudar no processo de delimitação e determinação dos segmentos e subsegmentos, bem como, o processo de segmentação textual. Assim, clarificamos que o processo de análise do elemento verbal apresenta possibilidades várias que não foram contempladas neste estudo.

3.5 Resumo do Capítulo

Neste capítulo, apresentamos os conceitos linguísticos que fornecem os fundamentos para a elaboração da metodologia para análise discursiva. Neste sentido, apresentamos :

1 – O conceito de texto e discurso, visto que, a nossa metodologia faz um percurso que se inicia com a análise completa da estrutura textual e sua conclusão e na apresentação da macroestrutura discursiva.

2 – A definição de proposição e a sua relação com os segmentos e subsegmentos.

3 – A definição terminológica para segmentos e subsegmentos em nossa metodologia.

4 – A classificação dos níveis textuais: micro, macro e superestrutural e a sua relação com a análise automática do discurso.

5 – A apresentação de elementos complementares que auxiliam na identificação dos segmentos e subsegmentos, bem como, seu papel na segmentação automática dos textos.

No próximo capítulo, apresentamos a metodologia desenvolvida no âmbito desta tese, a qual tem como objetivo realizar a análise discursiva e produzir automaticamente a macroestrutura/macroproposição de um texto sem a intervenção humana. Descrevemos precisamente cada um dos módulos que constituem a base metodológica, os quais representam cada uma das etapas a serem implementadas e executadas pelo *AuTema-Dis*.

Capítulo 4

Metodologia AuTema-Dis

4.1 A Proposta Metodológica

A proposta metodológica que apresentamos no âmbito desta tese foi edificada com o objetivo principal de realizar análise textual, direcionando os seus resultados à elaboração da macroestrutura/macroproposição discursiva. Para tal, desenvolvemos uma arquitetura do tipo modular em que a execução dos diferentes módulos é capaz de realizar a análise textual, considerando-se os diferentes níveis linguísticos que constituem na íntegra a estrutura do texto. A metodologia foi desenvolvida presumindo a sua futura demonstração e validação em um sistema computacional, denominado por *AuTema-Dis*, o qual encontra-se descrito no capítulo 5 desta tese.

No que diz respeito à constituição da metodologia, à elaboração da arquitetura foi prevista a realização de quatro etapas de análise distintas, mas relacionadas entre si. Cada módulo da arquitetura consiste em uma das etapas de análise e o resultado da execução de cada uma apresenta informações, para a execução da etapa seguinte até a conclusão de todo o processo. Assim, determinamos quatro módulos básicos para a realização da análise:

1. Módulo para identificação, classificação e segmentação dos constituintes textuais;
2. Módulo para organização arbórea –*DTS's* dos constituintes textuais;
3. Módulo para determinação das relações retóricas nas DTS's;
4. Módulo para identificação da macroproposição e produção da macroestrutura discursiva.

A edificação dos quatro módulos básicos que compõem a arquitetura foi realizada a partir de observações empíricas a respeito da constituição da estrutura textual. Acreditou-se que uma arquitetura modular estaria adequada ao tipo de análise que se deseja realizar, visto ser necessário garantir que cada um dos processos envolvidos pudesse ser realizado de forma autônoma. No entanto, apesar de se observar um determinado nível de independência entre os módulos, os resultados obtidos são necessariamente compartilhados nas sucessivas etapas que prosseguem à metodologia. Semelhante à constituição textual, a metodologia prevê um comprometimento entre todas as etapas de análise, não sendo possível se chegar a um resultado satisfatório se houver problema na realização de cada uma das etapas ou módulos de análise.

O modelo de arquitetura proposto para realizar a análise discursiva pode ser melhor identificado na figura 4.1.

4.2 Etapas da Metodologia - AuTema-Dis

4.2.1 Módulo 1 - Identificação e Segmentação dos Constituintes Textuais

A primeira etapa da metodologia tem como objetivo específico identificar as estruturas que constituem um texto e segmentá-las de acordo com a sua relação e importância com o tema do texto. Para tal, realizou-se, inicialmente, uma análise manual em um conjunto constituído por dez textos extraídos do corpus em formato digital do Jornal Público dos anos de 1994 e 1995, os quais fazem parte do conjunto *aprendizado* do corpus em Português Europeu *PE*, e encontram-se na seção de anexos desta tese.

O objetivo desta análise manual foi identificar nos textos selecionados um padrão composicional, no que se refere à categorização morfo-sintático-semântica. Buscou-se características que fossem passíveis de comporem um conjunto de regras, as quais pudessem servir de base a um sistema automático para análise textual. A partir dos resultados desta análise manual, no grupo de dez textos, verificou-se algumas divergências, as quais justificavam-se, em algumas situações, devido a questões subjetivas, próprias a cada analista, fato que não favorecia à padronização de regras. Da necessidade em criar uma padronização na maneira em como realizar a identificação das características, optou-se pela análise automática nos textos, para buscar a regularidade, o padrão.

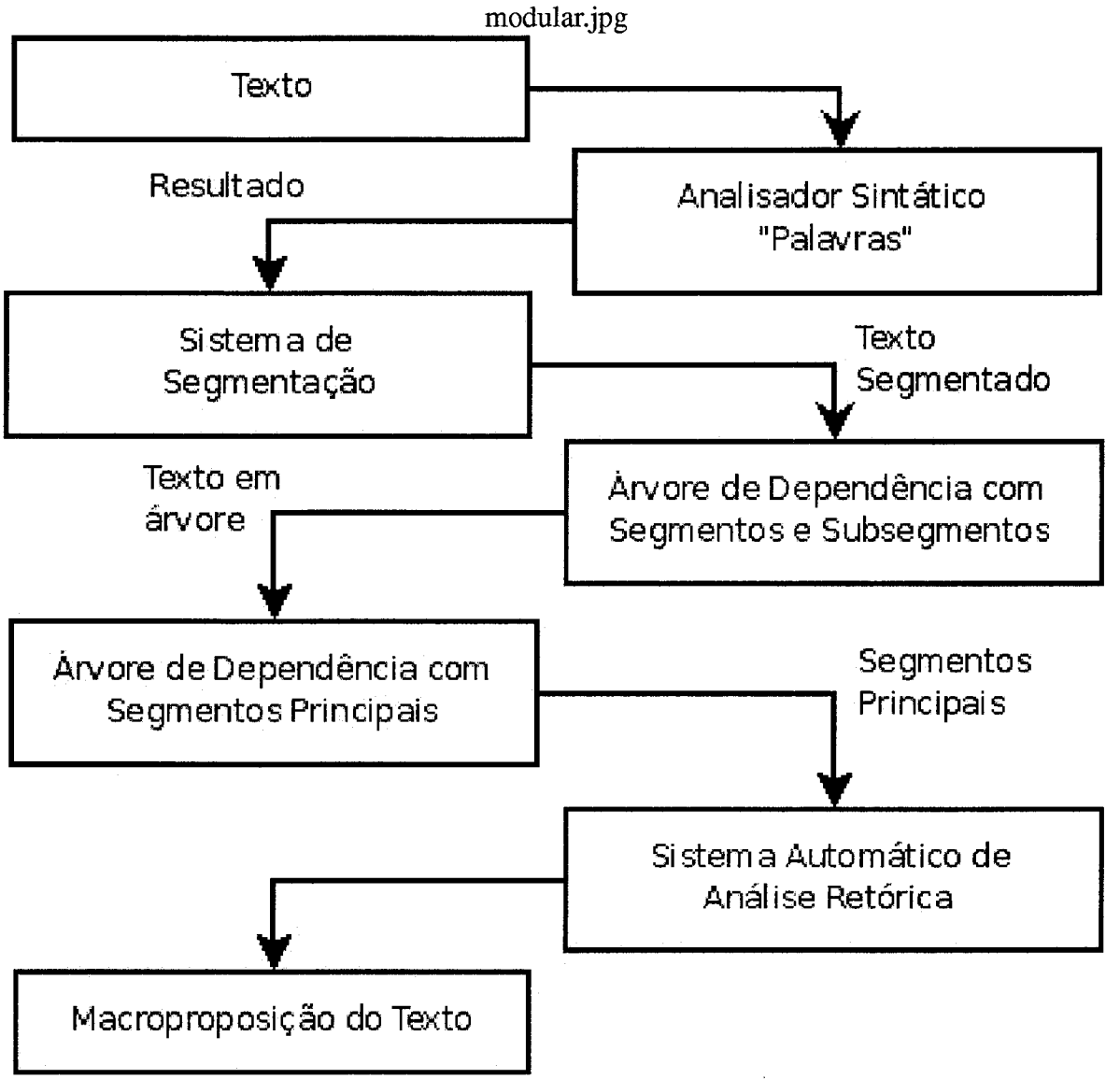

Figura 4.1: Arquitetura modular elaborada para análise textual – sistema AuTema-Dis.

4.2.1.1 Palavras - Analisador Automático

Determinados em obter um padrão específico na classificação das estruturas textuais, que propiciasse a elaboração de regras, as quais pudessem compor uma base heurística, selecionou-se um analisador sintático automático. O objetivo de fazer uso de um analisador automático é justamente solucionar o problema, referido anteriormente, isto é, gerir as diferentes possibilidades no processo de identificação e determinação dos constituintes textuais para uma mesma estrutura. Dos analisadores desenvolvidos para diferentes línguas, o que melhor apresentou resultados em comparação com a análise manual foi o *Palavras*, desenvolvido por Bick [1], no âmbito do projeto *VISL6*, no Institute of Language and Communication da University of Southern Denmark.

Além de apresentar resultados satisfatórios, o analisador *Palavras* exibe o resultado do processamento em estruturas arbóreas devidamente etiquetadas com a identificação morfo-sintática e, em alguns casos, uma notação semântica, em conformidade com a necessidade do utilizador previamente determinada. O analisador *Palavras* apresenta o resultado da análise morfo-sintático-semântico em uma codificação específica, a partir de uma gramática própria, desenvolvida especificamente para análise automática de textos, conforme apresentamos na figura 4.2 em que podem ser observadas as características morfo-sintático-semânticas presentes na saída da análise. No seguinte exemplo, apresentamos a figura 4.2 representativa da análise sintática do palavras.

```
STA:fcl
=SUBJ:np
==>N:art('o' <artd> F S)     A
==H:n('camara' F S) camara
=P:v-fin('nomear' PS 3S IND)  nomeou
=,
=ADVL:adv('entretanto' <kc>) entretanto
=,
=PRED:np
==>N:art('um' <arti> M S)    um
==H:n('grupo' <HH> M S) grupo
==N<:pp
===H:prp('de')  de
===P<:n('trabalho' <am> <act-d> M S) trabalho
=PRED:pp
==H:prp('com')  com
==P<:np
===>N:art('o' <artd> F S)    a
===H:n('finalidade' <am> F S) finalidade
===N<:pp
====H:prp('de') de
====P<:icl
=====P:v-inf('acompanhar' 0/1/3S) acompanhar
=====ADVL:adv('diariamente') diariamente
=====ACC:np
======>N:art('o' <artd> M S)     o
======H:n('abastecimento' <act> M S) abastecimento
(restante conteúdo omitido)
```

Figura 4.2: Exemplo parcial da saída do *Palavas* para o texto Jornal Público-19950726-079.

Como se pode ver na representação sintática do texto produzida pelo *Palavras*, cada palavra

é classificada na sua forma mais básica, havendo indicação de ordem morfológica associada à estrutura analisada. Além disso, o analisador atribui à classificação informações como o gênero, o número, o tipo de verbo e a sua conjugação úteis ao processamento automático do texto. A análise sintática é bastante linear e de fácil compreensão, sendo interpretada e compreendida de forma clara pelo analista que consegue reconhecer as relações entre os termos. Assim, optamos por utilizar o analisador *Palavras* como um recurso, isto é, uma ferramenta no módulo 1 da metodologia proposta. Ressaltamos que o processamento do *Palavras* não é um módulo de análise, mas sim uma ferramenta utilizada no âmbito do módulo 1.

No tocante à edificação das regras, os resultados da análise manual foram contrastados com as características identificadas a partir dos resultados das análises realizadas nos dez textos pelo *Palavras*. Os resultados e características da análise automática foram estudados minuciosamente e categorizados em uma base de dados. Obtivemos um conjunto inicial de características do cunho formal/conceitual, representativo das estruturas de dez textos em *PE*. A categorização serviu para a constituição das primeiras regras que compõem a primeira parte do módulo 1, desenvolvido para realizar a identificação automática das estruturas nos textos.

Conforme apresentamos, as primeiras regras foram identificadas a partir da análise manual e da análise automática realizada pelo *Palavras* em alguns textos selecionados a partir do corpus do Jornal Público. O conjunto das regras edificado tem como objetivo identificar, e segmentar as estruturas presentes nos textos, no tocante à categorização segmentos – subsegmentos, isto é, as unidades mínimas de significação. Abaixo, transcrevemos a sequência de atividades realizadas para elaboração das regras:

1. Realização da análise manual dos textos;
2. Realização da análise sintática automática com o *Palavras* nos textos;
3. Identificação das características da análise sintática (manual e automática) realizada nos textos selecionados;
4. Determinação dos limites de cada uma das estruturas dos textos analisados a partir do resultado do *Palavras*;
5. Elaboração das regras para identificação automática das estruturas textuais e posterior segmentação textual.

Da análise realizada, obtivemos uma aproximação do que poderia vir a constituir o conjunto de regras. Assim, compusemos a base inicial do sistema, que foi implementado em

linguagem prolog. Os dez textos iniciais foram submetidos ao sistema de segmentação implementado, apresentando, como era expectável, um resultado satisfatório, o que pode ser evidenciado na capítulo 5, na tabela 5.1, que apresenta a estatística e a avaliação das regras e do sistema. No entanto, foi necessário verificar se o conjunto sistematizado com regras obtidas pela análise manual e pela análise do *Palavras* poderia apresentar um resultado satisfatório no tocante à identificação e segmentação textual em um novo conjunto de textos em Português. As primeiras regras identificadas podem ser evidenciadas na Tabela 4.1.

Regras de Segmentação - Aprendizado 10 Textos - PE		
Regras Iniciais	N° de Ocorrências Manual	Identificação da Regra
N<:fcl	5	Informação Acessória /Complementar
Advl:pp	42	Circunstância Genérica
Advl:advp	7	Circunstância Genérica
Advl:fcl	1	Circunstância Genérica
Pred:pp	6	Circunstância Genérica
Advl:cu	1	Circunstância Genérica
App:prop	1	Circunstância Apositiva Nome Próprio
Co:conj-c('mas')	3	Oposição / Antítese
Pred:np	2	Elaboração - Circunstância Genérica
App:np	2	Informação Complementar Apositiva
Advl:adv - atemp	6	Quantificação Temporal
Advl:np ou Advl:n	1	Circunstância de Tempo Decorrido

Tabela 4.1: A tabela apresenta as regras iniciais identificadas a partir da análise dos 10 textos do Conjunto Aprendizado, em Português Europeu, do Jornal Público 1994/1995.

No tocante às regras demonstradas na tabela 4.1, observa-se que a classificação é feita mantendo-se a mesma referência obtida a partir do resultado do analisador *Palavras*. Além da classificação das regras, apresentamos o número de ocorrências em que cada uma delas foi identificada manualmente no conjunto aprendizado.

A nominalização das regras está relacionada à caracterização sintático-semântica. Assim, verifica-se, por exemplo, nas regras ADVL:pp, ADVL:advp, ADVL:fcl, Pred:pp, ADVL:cu que todas elas representam semanticamente a mesma característica, ou seja, as cinco regras fazem referência a um *circunstância genérica*. A determinação precisa da circunstância representada por cada uma das regras requer um estudo mais direcionado a este fim. Neste sentido, seria necessário e recomendável um estudo profundo e amplo, o que tornar-se-ia inviável no âmbito desta tese.

Para tornar legítimo o conjunto das regras, identificadas a partir do corpus *aprendizado*, foi necessário testá-las em outros textos, assim, selecionamos 40 novos textos, os quais fazem

parte dos *corpora* desta pesquisa. Os *corpora* são constituídos por dois corpus: um em Português Europeu – *PE* e outro em Português Brasileiro – *PB*, respectivamente, *Corpus A* e *Corpus B*, os quais encontram-se nos anexos desta tese.

No que se refere à constituição dos *corpora*, acreditamos ser importante uma breve apresentação, visto que, possibilitaram a identificação das regras iniciais, além de terem sido utilizados nos testes para a validação da metodologia a partir da implementação e execução em sistema automático.

4.2.1.2 Os Corpora

Optamos por formar os *corpora* com duas variantes do Português, *PEuropeu* e *PBrasileiro*, devido à diversidade na constituição estrutural e formal de ambas vertentes, identificadas em periódicos tipo jornal, os quais podem ser consultados no anexo 1 desta tese. Os textos são todos escritos, provenientes da comunicação no mundo real (língua em uso). Trata-se, especificamente, de um conjunto de dados lingüísticos reais, criteriosamente coletados; constituintes autênticos e representativos do Português nas suas variantes oficiais *PE* e *PB*. No que se refere ao texto jornalístico, a variedade tipológica, característica desse tipo de produção, atribui à investigação amplitude no que se refere à recolha dos dados, bem como, a possibilidade de criação de regras abrangentes, passíveis de serem aplicadas em uma extensiva variedade de produções textuais.

Os *corpora* foram caracterizados manualmente antes de serem processados pelo *Palavras*, foram realizadas análises morfo-sintáticas e algumas marcações semânticas. Além desta caracterização formal/conceitual, foi possível atribuir algumas informações relativas a sua constituição estrutural, isto é, número de palavras, linhas e parágrafos identificados em cada um dos textos; e o número de estruturas, sendo essas reconhecidas, neste trabalho, ao início de uma frase ou oração e finalizada com um ponto. A categorização foi realizada para verificar se os textos selecionados apresentavam uma harmonia estrutural, para que a sua eleição não gerasse inconsistência e grandes diferenças na determinação do número de estruturas que os compõem.

Acreditamos, na altura da composição dos *corpora*, que se os textos apresentassem grandes diferenças estruturais, essas poderiam desfavorecer a estatística relativa à identificação dos constituintes do texto pelo sistema computacional. Assim, procuramos restringir o tamanho dos textos, isto é, buscamos um padrão estrutural em prol de obtermos um desempenho mais

adequado e correto do sistema implementado para os testes, isto é, garantir uma melhor performance do AuTema-Dis. A fim de visualizar mais claramente as características estruturais apresentadas nos *corpora* que constituem esse estudo, apresentamos as tabelas 4.2, 4.3 e 4.4, com a classificação estrutural dos textos.

Corpus Aprendizado Corpus Português Europeu: Jornal Público 1994 e 1995 Características do corpus				
Textos	Palavras	Linhas	Estruturas	Parágrafos
Nº 19940101-007	92	10	5	2
Nº 19941012-035	54	7	2	2
Nº 19941214-076	143	15	7	3
Nº 19950519-057	96	10	3	2
Nº 19950725-025	59	7	2	3
Nº 19950726-079	120	13	3	4
Nº 19950916-121	73	8	4	4
Nº 19950916-157	122	13	6	3
Nº 19950917-041	105	13	4	3
Nº 19950814-011	84	9	4	2
Média	94.8	10.5	4	2.8

Tabela 4.2: A tabela apresenta as características estruturais do conjunto aprendizado/treino.

Constituídos os *corpora* e justificadas as suas escolhas, iniciaram-se os testes na totalidade dos textos. As regras, com as quais os novos textos foram processados, foram elaboradas inicialmente com as características identificadas em apenas dez textos do jornal Público. A partir do processamento no restante textos dos *corpora*, identificou-se que seria necessário propor novas regras, pois os *corpora* analisados apresentaram novas características, as quais não haviam sido contempladas na análise realizada nos dez textos em *PE*. Assim, foi necessário reestruturar o conjunto, incluindo algumas novas regras, a fim de que o sistema realizasse adequadamente a identificação e segmentação dos textos.

A partir do resultado da análise realizada na totalidade dos *corpora*, foi possível identificar novas regras, no total de três, as quais foram agregadas ao conjunto inicial, e que podem ser evidenciadas abaixo na tabela 4.5, juntamente com o número de vezes que cada regra aparece nos textos dos *corpora*.

Corpus Avaliação/Teste Corpus Português Europeu: Jornal Público 1994 e 1995 Características do corpus				
Textos	Palavras	Linhas	Estruturas	Parágrafos
N° 19940504-070	99	9	4	2
N° 19940505-024	69	7	5	2
N° 19940505-071	167	16	5	4
N° 19941911-083	169	16	6	6
N° 19941012-011	137	12	6	2
N° 19941025-045	211	18	6	4
N° 19950416-032	161	15	4	2
N° 19950795-167	143	13	6	2
N° 19950912-022	99	9	4	2
N° 19950924-121	177	15	9	2
N° 19950422-141	78	7	4	2
N° 19950423-011	113	9	4	2
N° 19950629-083	123	12	5	3
N° 19950629-119	123	10	5	3
N° 19951011-139	93	8	3	1
N° 19951011-150	115	10	8	4
N° 19951114-163	100	8	3	1
N° 19951114-169	121	10	6	1
N° 19951220-045	84	7	3	2
N° 19951229-044	86	8	3	3
Média	123.4	10.95	4.95	2.5

Tabela 4.3: A tabela apresenta as características estruturais do conjunto avaliação/teste em Português Europeu.

4.2.1.3 Conjunto de regras para a segmentação

Conforme mencionamos, o resultado da análise manual e análise do *Palavras* nos textos dos *corpora*, bem como, o tratamento/manipulação deste resultado revelou que os traços e as características lingüísticas não ocorrem de forma aleatória, sendo possível categorizar e quantificar regularidades (padrões) em forma de regras. As regras, devidamente implementadas em prolog, são utilizadas para identificação dos constituintes textuais, classificação com segmentos e subsegmentos e para a delimitação dos seus limites e as suas fronteiras nas estruturas das quais fazem parte nos textos. De uma forma mais específica, as regras propostas são utilizadas para identificar os constituintes dos textos; de uma forma mais ampla as regras podem também auxiliar no processo de segmentação textual determinando os locais em que um texto pode ser segmentado, em conformidade com os critérios estabelecidos a partir das análises manuais e automáticas.

Corpus Avaliação/Teste Corpus Português Brasileiro: Folha de São Paulo 1994 e 1995 Características do corpus				
Textos	Palavras	Linhas	Estruturas	Parágrafos
Nº FSP950101-011	125	11	4	1
Nº FSP950101-032	91	9	4	2
Nº FSP950101-054	63	5	6	1
Nº FSP950101-084	117	15	6	7
Nº FSP950111-014	82	8	4	2
Nº FSP950111-026	132	15	6	4
Nº FSP950111-034	111	13	8	4
Nº FSP950111-036	59	5	2	1
Nº FSP950117-048	178	18	9	4
Nº FSP950117-074	119	13	6	5
Nº FSP940101-132	62	6	5	2
Nº FSP940101-124	139	13	8	4
Nº FSP940101-107	74	7	9	3
Nº FSP940101-102	175	16	9	4
Nº FSP940101-095	262	23	19	8
Nº FSP940101-092	164	17	7	4
Nº FSP940101-085	131	12	5	4
Nº FSP940101-079	113	11	3	3
Nº FSP940101-074	137	12	9	4
Nº FSP940101-066	128	12	8	4
Média	123.1	12.05	6.85	3.55

Tabela 4.4: A tabela apresenta as características estruturais do conjunto avaliação/teste em Português Brasileiro.

Na sequência, apresentamos o conjunto definitivo das regras para reconhecimento e a classificação dos segmentos e subsegmentos, conforme as características identificadas na totalidade dos *corpora*, como pode ser identificado nas tabelas 4.6 e 4.7. Observamos, outrossim, que há regras distintas para identificação dos segmentos e dos subsegmentos. Como previámos, o número de regras para a identificação dos segmentos apresenta um número mais reduzido do que as regras para os subsegmentos. Acreditamos que essa diferença no número de regras esteja relacionada à complementação verbal suportada em Língua Portuguesa.

Como observamos na figura 4.2, o analisador *Palavras* analisa e atribui uma classificação em todas as palavras do texto analisado. As caracterísitcas são também atribuídas às estruturas, isto é, frases ou orações, e são essas características, atribuídas às estruturas com a sua específica codificação que é utilizada em comunhão ao conjunto de regras proposto em nossa metodologia. A associação das características atribuídas às estruturas textuais pelo *Palavras* mais as regras propostas para identificação dos constituintes podem determinar em que ponto

Regras na Totalidade dos *Corpora* 50 textos	N° de Ocorrências Manual	N° de Ocorrências Sistema
N<:fcl	38	40
Advl:pp	239	205
Advl:advp	15	16
Advl:fcl	30	34
Pred:pp	19	16
Advl:cu	7	7
App:prop	16	14
Advl:acl	8	8
Sta:icl	10	13
Co:conj-c('mas')	14	12
Pred:np;	13	7
App:np	8	10
Advl:adv - atemp	52	47
Advl:adv - aloc	1	3
Advl:np ou Advl:n	6	8

Tabela 4.5: A tabela apresenta o número de ocorrências de cada uma das regras na totalidade dos *corpora*, considerando-se a identificação manual e automática.

Regra para Segmento	Identificação da Regra
UTT:acl	Enunciado sem verbo - títulos, manchetes (jornal) e cabeçalhos
EXC:fcl	Estrutura Exclamativa
QUE:fcl	Estrutura Interrogativa
NPHR:prop	Enunciado sem verbo - com estrutura nominal própria (nome próprio)
NPHR:np	Enunciado nominal
STA:fcl	Enunciado com oração finita

Tabela 4.6: A tabela apresenta as regras para a identificação dos segmentos, bem como, a sua definição terminológica.

a estrutura textual oferece a possibilidade mais adequada para a segmentação, bem como, a determinação das suas fronteiras.

Na sequência deste capítulo, apresentamos a constituição do 2° módulo proposto na metodologia, o qual utiliza as regras para identificação dos segmentos e subsegmentos, propostas no 1° módulo, para constituir a etapa responsável pela organização destes constituintes em árvores DTS's e a sua futura automatização.

Regras para Subsegmentos	Identificação da Regra
N<:fcl	Informação Acessória /Complementar
Advl:pp	Circunstância Genérica
Advl:advp	Circunstância Genérica
Advl:fcl	Circunstância Genérica
Pred:pp	Circunstância Genérica
Advl:cu	Circunstância Genérica
App:prop	Circunstância Apositiva Nome Próprio
Advl:acl	Avaliação
Sta:icl	Ação
Co:conj-c ('mas')	Oposição / Antítese
Pred:np	Elaboração - Circunstância Genérica
App:np	Complementação Nominal Apositiva
Advl:adv - atemp	Quantificação Temporal
Advl:adv - aloc	Quantificação Locativa
Advl:np ou Advl:n	Circunstância de Tempo Decorrido

Tabela 4.7: A tabela apresenta regras para identificação dos subsegmentos, bem como, a sua classificação terminológica.

4.2.2 Módulo 2 - Organização Arbórea – DTS's

O segundo módulo da metodologia prevê a organização dos constituintes textuais, identificados no módulo 1, em árvores tipo *DTS's* e a sua futura automatização. O conceito **DTS's** foi desenvolvido no âmbito deste trabalho e, conforme apresentamos no capítulo 3, as DTS's diferenciam-se de outras representações arbóreas por associarem não apenas características sintáticas. Trata-se de uma organização estrutural arbórea idealizada a partir da interação de características linguísticas (morfo-sintático-conceituais), direcionada a representar automaticamente o texto hierarquicamente estruturado. As árvores de dependência de segmentos são utilizadas, neste estudo, com o objetivo demonstrar a hierarquia entre os segmentos que compoõem a estrutura textual a partir da interação de características estruturais, sintáticas e conceituais procedentes do resultado da análise manual e automática nos textos dos *corpora*, conforme apresentamos na sequência desta seção.

No que se refere às características estruturais ponderadas para a organização dos constituintes textuais nas árvores, essas são identificadas tendo em conta as informações advindas do módulo 1, isto é, identificação dos segmentos/subsegmentos e segmentação das estruturas. Conforme apresentamos, as características que determinam as possibilidades para a segmentação são constituídas a partir da análise sintática realizada pelo *Palavras*. O analisador

apresenta um resultado codificado, conforme apresentamos na figura 4.2, em que é possível observar a análise das estruturas de um texto, devidamente codificadas/etiquetadas. Um fator relevante a ser considerado nessa etapa da análise é que, além da codificação e da etiquetagem das estruturas, o resultado do *Palavras* provê o nível de profundidade em que cada um dos constituintes se encontra no interior das estruturas das quais fazem parte.

A categorização dos níveis de profundidade em que se encontram os segmentos, que compõem as estruturas ao longo do texto, é relevante para determinar a composição e a organização hierárquica das árvores DTS's. A classificação dos níveis de profundidade apresentados pela análise realizada pelo *Palavras* fornece os dados que podem ser incorporados às regras de segmentação. Os níveis de profundidade apresentados a partir do *Palavras* podem ser observados abaixo na figura 4.3.

A marcação dos níveis presente no resultado da análise automática do *Palavras*, associada às regras de segmentação proposta no módulo 1, determina a disposição estrutural dos segmentos nas árvores DTS's. Em conformidade com o nível de profundidade que os constituintes ocupam nas estruturas, é possível identificar, em algumas situações, o tipo de relação que se estabelece entre eles, seja ela, *hipotaxis* ou *parataxis*[1]. Desta forma, os constituintes identificados automaticamente nas árvores DTS's são organizados a partir da interação entre as características sintático-estruturais, e as relações, sejam elas hipotáticas ou paratáticas, entre os constituintes possibilita-nos classificá-los de acordo com o papel que desempenham na estrutura, isto é, se o constituinte desempenha o papel de segmento ou de subsegmento, definições apresentadas no capítulo 3.

No que se refere às características conceituais, envolvidas no processo de automatização das DTS´s, retomamos o que foi apresentado no capítulo 3, em que relacionamos a identificação dos segmentos e subsegmentos à representação das proposições que compõem uma determinada estrutura discursiva. Conforme discutiu-se, as proposições são abstrações que o autor explicita a partir das relações entre segmentos e subsegmentos que compõem a estrutura textual; eles são a contrapartida linguística de uma proposição. Assim, se as relações entre os segmentos e subsegmentos representam as proposições, esses constituintes, ao serem organizados em DTS's, estão organizando conceitos, idéias enquanto concretas ocorrências.

No entanto, vale ressaltar que, no caso da metodologia que propomos para realizar a sistematização e automatização na identificação dos segmentos, a relação *segmento = proposição* nem sempre ocorre de forma linear/direta. Esta particularidade foi evidenciada em nossos resultados, devido a forma como as regras para identificação dos segmentos foram consti-

[1]Hipotaxis e Parataxis - O conceito de parataxis aqui é entendido como a coordenação entre as estruturas nos textos, com ou sem o uso de elementos coordenadores. A hipotaxis, por sua vez, apresenta a relação de dependência ou subordinação entre as estruturas nos textos, podendo ser identificada uma hierarquia entre os elementos.

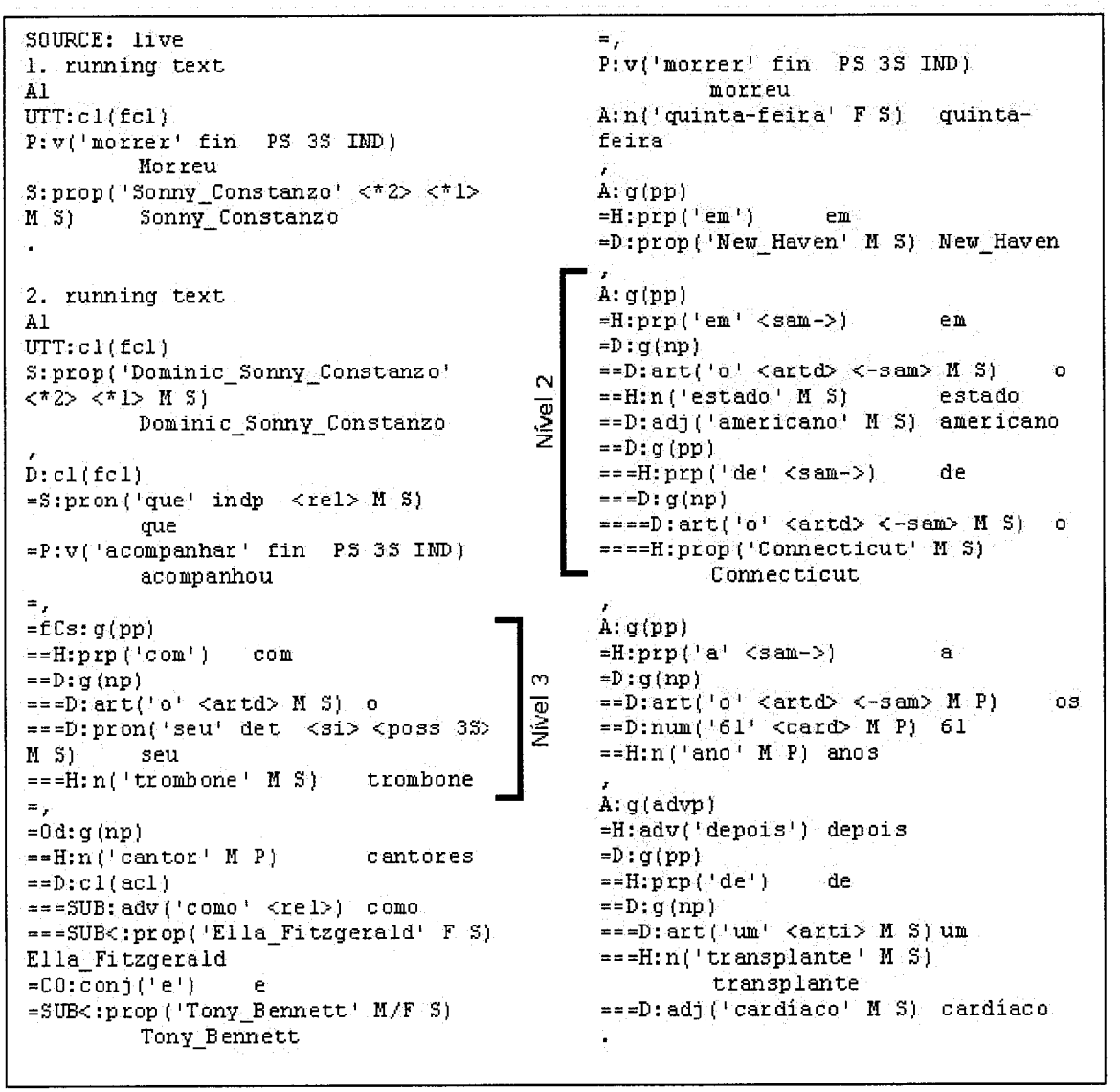

Figura 4.3: A figura apresenta a análise automática do *Palavras* com a marcação dos níveis de profundidade em que se encontram os constituintes na estrutura – texto publico-19950726-079.

tuídas, isto é, priorizando, em alguns casos, padrões formais e estruturais às características conceituais. Outro fato que justifica essa *não-linearidade* é que as regras são utilizadas a partir da execução em um sistema automático sem interferência humana, o qual identifica os segmentos, realiza a segmentação e os organiza em árvores conforme o seu papel na estrutura textual. A interação entre as características sintáticas e estruturais condicionam e estão condicionados por questões conceituais no que se refere à composição automática das DTS's. No caso desta investigação, a disposição tipo DTS's organiza os segmentos e os subsegmentos hierarquicamente, conforme papel que representam na estrutura textual não sendo apenas uma representação sintática, como a maioria das representações em árvores.

A disposição dos constituintes textuais nas DTS's estão condicionadas ao nível de profundidade que ocupam na estrutura da qual fazem parte, bem como, a relação que se estabele entre eles, visto que o segmento detém um papel subordinante em relação aos subsegmentos. As DTS's organizam os segmentos principais como "nós"do nível 1 da árvore e os subsegmentos são identificados como "nós"de níveis 2 e 3, conforme figura 4.4. Os demais subsegmentos, ou seja, aqueles que se encontram em níveis de profundidade posterior ao 3° nível, isto é, subsegmentos do 4° e 5° níveis, não são organizados de maneira hierárquica em "nós", são mantidos em "nós"únicos e indissociáveis, conforme apresentado na figura 4.4.

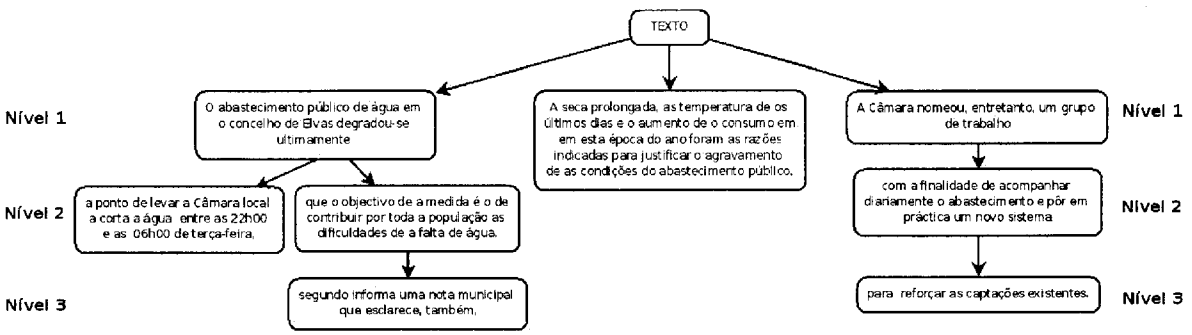

Figura 4.4: A figura representativa da organização hierárquica de um texto em uma DTS com especificação dos níveis – texto publico-19950726-079 .

4.2.2.1 Regras de Segmentação Textual e Identificação dos Níveis dos Constituintes

O objetivo deste módulo da metodologia é organizar os segmentos e os subsegmentos em árvores do tipo DTS's, conforme a representatividade e o compromisso que cada uma das estruturas desempenha em relação ao tema do texto. Para realizar tal organização, a metodologia prevê a utilização das regras de segmentação, elaboradas para a identificação dos segmentos no módulo 1, e de informação complementar relacionada ao nível de profundidade em que se encontram cada um dos constituintes no texto.

Em uma primeira análise, acreditou-se que as regras de segmentação poderiam ser suficientes para organização dos segmentos nas árvores, no entanto, essa não foi a realidade observada na prática. As regras de segmentação prestavam-se para a identificação linear dos segmentos, mas não os distiguia entre si. Assim, foi necessário associar às regras uma característica específica que, a partir da comunhão de ambas, pudesse classificar os segmentos em função do seu papel na estrutura do texto, visto ser evidente que algumas estruturas detém um caráter mais subordinante do que outras. Considerando que os constituintes textuais desempenham

diferenciados papéis na estruturação de um texto, recorremos à análise do *Palavras* na busca de características que pudessem ser associadas às regras e que contemplassem essa diferença entre os segmentos. Desta forma, evidenciamos que o resultado da análise apresentava, para cada um dos segmentos, a marcação de nível de profundidade na estrutura, característica que convém ao propósito da estruturação arbórea.

Definimos na metodologia que as regras de segmentação recebem a caracterização de nível de profundidade, isto é, cada regra que identifica um segmento ou um subsegmento traz agregada uma informação que determina o lugar que cada segmento ou subsegmento pode ocupar na árvore DTS, conforme evidenciamos na tabela 4.8. *A priori*, os constituintes identificados como segmentos ocupam os *nós* de 1° nível; já os constituintes identificados como subsegmentos ocupam os *nós* de 2° e 3° níveis. Nota-se que a análise realizada pelo *Palavras* determina níveis de profundidade além do 2° e 3°, entretanto, para fins metodológicos e de implementação do sistema, optou-se em considerar para segmentação e estruturação arbórea apenas estes dois níveis, ficando dos demais níveis associados aos *nós* de nível 2 e 3.

Níveis de Profundidade	Regra para Segmento
Todos segmentos encontram-se no nível 1 de profundidade na estrutura	UTT : acl EXC : fcl QUE : fcl NPHR : prop NPHR : np STA : fcl
Níveis de Profundidade	Regras para Subsegmentos
Todos segmentos encontram-se nos níveis 2 e 3 de profundidade na estrutura	N<:fcl Advl:pp Advl:advp Advl:fcl Pred:pp Advl:cu App:prop Advl:acl Sta:icl Co:conj-c ('mas') Pred:np App:np Advl:adv - atemp Advl:adv - aloc Advl:np ou Advl:n

Tabela 4.8: A tabela representa as regras de segmentação dos constituintes textuais e os níveis de profundidade que os segmentos e os subsegmentos podem ocupar em uma estrutura.

A importância desta etapa na metodologia está relacionada à formalização e à hierarquia das

estruturas que constituem o texto, e o seu papel em relação ao tema do discurso. O módulo 2 da metodologia foi proposto com o objetivo realizar automaticamente a organização dos segmentos e dos subsegmentos, identificados e segmentados no módulo 1, nas DTS's. A organização DTS propicia a identificação de como as proposições, representadas através dos segmentos e subsegmentos dispostos dispostss na estrutura textual, organizam-se para comporem o tema do texto.

Na próxima seção apresentamos o módulo 3 da metodologia, o qual propõe identificar automaticamente algumas relações retóricas entre os segmentos organizados nas DTS's.

4.2.3 Módulo 3 - Identificação das Relações Retóricas em DTS´s

O módulo 3 apresenta a metodologia para a identificação automática das relações retóricas entre segmentos e subsegmentos a partir da configuração em DTS's. A constituição deste módulo conta com as informações advindas dos módulos 1 e 2. O módulo 1 identifica e segmenta os constituintes do texto; o módulo 2 recebe e processa as informações do módulo 1, classifica e organiza os constituintes em árvores de dependência. A classificação é realizada de acordo com nível de profundidade em que os constituintes se encontram na estrutura da qual fazem parte e conforme o papel que cada um desses elementos representa na constituição do tema. O módulo 3 utiliza a organização arbórea realizada no módulo 2 e atribui algumas relações retóricas entre os segmentos, *nós* de 1° nível e subsegmentos, *nós* de 2° e 3° níveis.

As árvores DTS's apresentam o texto organizado em uma estrutura esquemática na qual os segmentos e subsegmentos aparecem hierarquicamente dispostos. Este tipo de organização é favorável à identificação das relações retóricas, visto que, as relações manifestam-se entre os constituintes respeitando a ordenação e a subordinação existente entre esses elementos, aliás, as próprias relações condicionam e são condicionadas pela organização estrutural do texto. Assim, optou-se por considerar a distribuição dos segmentos em árvores DTS's para favorecer a identificação das relações retóricas, tendo em vista a sua posterior automatização.

O primeiro passo para a elaboração da metodologia, que constitui o módulo 3, foi identificar características presentes nos textos dos *corpora* que pudessem estar relacionadas diretamente à representação das relações retóricas. Recorreu-se aos textos dos *corpora* previamente analisados pelo *Palavras*, para identificar padrões linguísticos da ordem morfo-sintático-semântica que pudessem determinar uma relação retórica, numa relação *um-para-um*, isto

é, uma característica equivale/representa uma relação. Além disso, a busca pelos padrões que deveriam contemplar especificidades passíveis de serem implementadas computacionalmente, visto que, a metodologia prevê, na sua origem, que todos os módulos sejam implementados e executados em sistema computacional.

Concomitante à busca dos padrões linguísticos que pudessem indicar para determinadas relações retóricas, foi-se realizando manualmente nos textos a identificação de relações retóricas entre os seus constituintes. A atribuição manual das relações entre as estruturas dos *corpora* proporcionou a aquisição de uma maior sensibilidade com os dados e características relacionadas à organização estrutural dos textos, à organização dos segmentos e subsegmentos e à composição/constituição do tema do texto. Como resultado desta atividade, foi possível verificar que a melhor forma de identificar as relações retóricas considerando-se apenas informações linguísticas da superfície textual, seria associar as relações retóricas às regras de segmentação com marcação de nível, criadas para o módulo 2. Desta forma, foi possível excluir-se toda e qualquer tipo de intervenção humana, já idealizando a implementação do sistema computacional AuTema-Dis, projetado para testar e representar a metodologia.

Reconhecemos que a avaliação manual possibilita a atribuição de uma maior variedade de relações retóricas, pois conta com a sensibilidade, a atenção e adequação do analista, quanto às possibilidades nas combinações das estruturas, o que não pode ser considerado, por exemplo, na execução da mesma atividade por um sistema computacional. Para a realização do processo de análise manual, utilizou-se dois conjuntos de relações retóricas, o conjunto proposto por Mann e Thompson [24] e o conjunto proposto Marcu [2], ambos grupos encontram-se nos anexos desta tese. Apesar de contarmos com dois grupos de relações, as relações retóricas proposta por ambos autores não foram utilizadas na íntegra em nosso estudo, optamos pela utilização de algumas das relações de Mann e Thompson e de Marcu de acordo com a necessidade de caracterizarmos as relações identificadas nos textos dos *corpora*. Das duas propostas, foram utilizadas as seguintes relações retóricas:

- Mann and Thompson - Relações Retóricas: circunstância; Avaliação; Antítese, Elaboração.
- Daniel Marcu - Relações Estruturais: Same-unit; Parentética (a qual não foi considerada para a automatização, foi utilizada somente no processo manual).

No âmbito desta investigação, além das relações selecionadas dos dois grupos mencionados, foi necessário identificar um novo grupo de relações, desenvolvidas exclusivamente para suprir particularidades evidenciadas nos textos dos *corpora*. As novas relações foram

definidas em conformidade com a teoria das relações retóricas que prevê na sua origem a possibilidade de livre ampliação, de acordo com as particularidades de cada texto, com os critérios e objetivos estabelecidos pelo analista e com a finalidade da análise. Neste sentido, além das seis relações retóricas utilizadas, provenientes do rol das relações de Mann e Thompson e Marcu, agregou-se mais cinco relações, são elas: Apositiva de Nome Próprio; Quantificação Temporal, Quantificação Locativa, Ação, Circunstância de Tempo Decorrido. A figura 4.9 apresenta detalhadamente as definições para as novas relações introduzidas, seguindo a linha proposta pela RST.

Nova Relação Retórica	Regras de Segmentação	Restrições e Efeitos	Exemplo
Apositiva de Nome Próprio	App:prop	**Núcleo**: apresenta uma informação nominal pouco específica. **Satélite**: apresenta um Nome Próprio que especifica a informação descrita no núcleo. **Restrições N+S**: o S especifica através de uma expressão nominal própria, relacionada nominalmente ao que foi apresentado pelo N. **Efeito**: o leitor recebe do S a especificação através de um nome próprio daquilo que é mostrado no N.	(...) a posse dos presidentes do Banco do Brasil, **Paulo César Ximenes**, e da Caixa Econômica Federal, **Sérgio Cutolo**. FSP950111-034
Circunstância de Tempo ou Quantificadora	Advl:adv atemp	**Núcleo**: não há. **Satélite**: apresenta um elemento temporal. **Restrições N+S**: o S apresenta uma característica temporal para a situação descrita no N. **Efeito**: o leitor reconhece quando o fato apresentado pelo N foi realizado.	A inauguração do Mercado Abastecedor de Coimbra (MAC) foi **ontem**, ao fim da tarde, interrompida. Público19950705-167
Circunstância de Lugar ou Quantificadora	Advl:adv aloc	**Núcleo**: não há. **Satélite**: apresenta um elemento lugar. **Restrições N+S**: o S apresenta uma característica locativa para a situação descrita no N. **Efeito**: o leitor reconhece onde o fato apresentado pelo N foi realizado.	(...) Certamente passarei **aqui** a maior parte de janeiro, (...) FSP950101-084

Nova Relação Retórica	Regras de Segmentação	Restrições e Efeitos	Exemplo
Ação	Sta:icl	**Núcleo**: apresenta uma situação. **Satélite**: demonstra uma ação a ser realizada a partir do que é apresentado pelo N. **Restrições N+S**: a situação apresentada pelo N condiciona a ação descrita pelo S. **Efeito**: o leitor percebe que o S apresenta uma ação realizada ou a ser realizada condicionada pelo que é descrito no N.	(...) a Alemanha deu um passeio, **vencendo fácil com a charmosa dupla Steffi Graf e Michael Stich.** FSP940101-095
Temporal / Tempo decorrido	Advl:np ou Advl:n	**Núcleo**: não há. **Satélite**: caracteriza uma situação temporal concluída ou em andamento com base no que é apresentado no N. **Restrições N+S**: S situa no tempo (concluído ou não) a informação apresentada no N. **Efeito**: o leitor reconhece a temporalidade do que é descrito no N.	**A certa altura**, uma mulher das Caxinas, que já tinha boné, não estava disposta a deixar Gomes, (...) Publico19950924-121

Tabela 4.9: A tabela representa as novas relações retóricas desenvolvidas no âmbito desta investigação, evidenciadas na totalidade dos *corpora*.

Na sequência, após ter sido determinado manualmente o grupo de relações retóricas mais evidenciadas nos textos dos *corpora*, iniciou-se, efetivamente, a composição da parte da metodologia que propõe automatizar a identificação das relações retóricas entre os segmentos e os subsegmentos nos textos. Conforme mencionamos acima, procurou-se características na superfície textual que pudessem estar associadas a uma determinada relação na ordem de *um-para-um*.

Tendo em vista o tipo de investigação desenvolvida, determinamos que a melhor opção para a automatização das relações seria associá-las a uma regra de identificação dos segmentos. No módulo 1, foram produzidas 16 regras para identificação dos **subsegmentos** e 6 regras para identificação dos **segmentos**, todavia, somente as regras para a identificação dos subsegmentos é que estão indexadas uma relação retórica, as demais, como é o caso das regras para identificação dos segmentos, estas não apresentam nenhuma relação retórica a elas associadas.

Verificou-se que mais de uma regra para a identificação dos subsegmentos pode estar inde-

xada um mesmo tipo de relação como é o caso das regras: ADVL:pp; ADVL:fcl; ADVL:advp; ADVL:cu e Pred:pp todas estão indexadas à relação retórica *circunstância genérica*. No rol das relações utilizadas na metodologia existem 15 regras para identificação dos subsegmentos implementadas, e somente 11 relações retóricas a elas indexadas. No caso das 6 regras para identificação dos segmentos, optou-se por não indexar nenhuma relação retórica a elas, em conformidade com as diretrizes adotadas previamente para a edificação desta metodologia investigação.

4.2.3.1 O conjunto das relações retóricas na metodologia AuTema-Dis

No módulo 3, propusemos a metodologia para automatização das relações retóricas entre os constituintes do texto. Conforme apresentamos, as relações retóricas elegidas, que constituem o conjunto a ser avaliado através da implementação e execução no sistema AuTema-Dis, perfazem um total de 11 relações, sendo estas relações apenas para as ligações entre os segmentos e subsegmentos, e subsegmentos – subsegmentos. No momento, a metodologia ainda não está desenvolvida para contemplar as relações entre dois ou mais segmentos e alguns casos de relações entre segmentos e subsegmentos, o que justifica o número reduzido de relações implementadas no sistema, conforme podemos identifica na tabela 4.10.

Relações Retóricas - Metodologia AuTema-Dis
Same-Unit
Circunstância Genérica
Circunstância Apositiva Nome Próprio
Avaliação
Ação
Oposição/Antítese
Elaboração - Circunstância Genérica
Complementação Nominal Apositiva
Quantificação Temporal
Quantificação Locativa
Circunstância de Tempo Decorrido

Tabela 4.10: A tabela apresenta as 11 relações retóricas que constituem o sistema AuTema-Dis.

Ao estruturarmos a metodologia, optamos por identificar apenas algumas relações retóricas, devido à complexidade no processo de sistematização em manipular informações semânticas inerentes às relações. Outrossim, observamos que as relações imputam referências conceituais que, em algumas situações, não podem ou não estão delimitadas por nenhum

constituinte linguístico de superfície, do tipo marcador discursivo, o qual poderia indicar ou determinar de forma mais evidente a sua significação. Outro ponto a ser mencionado é o caso de algumas relações que, apesar de terem sido evidenciadas na análise manual, não fazem parte do conjunto utilizado em nossa investigação, como é o caso da *relação estrutural parentética*, proposta por Daniel Marcu. A relação mencionada foi evidenciada algumas vezes na análise dos *corpora*, todavia devido a questões relacionadas ao processo de implementação, optamos por não incluí-la no rol das relações que são utilizadas neste estudo.

Conforme apresentamos e exemplificamos, as relações retóricas as já existentes e as que foram desenvolvidas no âmbito desta investigação estão indexadas às regras para identificação dos segmentos e dos subsegmentos e podem ser evidenciadas na tabela 4.11. Relacionou-se cada uma das relações a uma regra de subsegmentação.

Relações Retóricas	Regra para Segmento
A definir	UTT : acl EXC : fcl QUE : fcl NPHR : prop NPHR : np STA : fcl
Relações Retóricas	Regra para Subsegmento
Same-Unit	N<:fcl
Circunstância Genérica	Advl:pp Advl:advp Advl:fcl Pred:pp Advl:cu
Circunstância Apositiva Nome Próprio	pp:prop
Avaliação	Advl:acl
Ação	Sta:icl
Oposição/Antítese	Co:conj-c ('mas')
Circunstância Apositiva Nome Próprio	Pred:np
Complementação Nominal Apositiva	App:np
Quantificação Temporal	Advl:adv - atemp
Quantificação Locativa	Advl:adv - aloc
Circunstância de Tempo Decorrido	Advl:np ou Advl:n

Tabela 4.11: A tabela representa as relações retóricas indexadas às regras para identificação dos segmentos e subsegmentos.

No que ser refere à composição da metodologia, especificamente à parte relacionada ao processo de automatização das relações, optamos por não atribuir relações retóricas entre os constituintes do tipo *segmentos*. Apesar de ter sido elaborado um conjunto completo de re-

gras para a identificação e segmentação de todos constituintes presentes na estrutura textual, no caso dos constituintes identificados como *segmentos*, isto é, estruturas que ocupam os *nós* de nível 1.

O processo de investigação realizado até o presente momento não possibilita refinarmos a metodologia ao ponto de suportar as relações entre segmentos, além de não termos considerado esse um dos nossos objetivos na elaboração da pesquisa. No entanto, o estudo realizado revelou a necessidade desenvolver análises específicas para a atribuição de relações entre esses *segmentos*, o que não pode ser realizado neste estudo devido à amplitude deste tipo de análise.

No tocante à identificação das relações retóricas,ressaltamos a sua importância para a concretização da etapa seguinte, isto é, identificação da macroproposição e a sua representação através da macroestrutura, conforme apresentamos próxima seção, seqüência deste capítulo.

4.2.4 Módulo 4 - Representação Estrutural da Macroproposição Textual

A metodologia que constitui o módulo 4 foi construída a partir dos resultados obtidos pela realização dos três módulos anteriores, propostos para esta investigação. Trata-se de um *módulo-resultado* em que os constituintes textuais encontram-se organizados linearmente em uma macroestrutura, representativa das macroproposições que se configuram ao longo da estrutura analisada em cumprimento ao tema tratado no texto.

Revisando a metodologia que constitui a base do que vem a ser o sistema AuTema-Dis, salienta-se que a organização metodológica encontra-se articulada em todos os níveis propostos para realização da análise textual. Dessa forma, tornam-se inicialmente e necessários e apropriados os resultados do primeiro módulo, obtidos a partir da aplicação das regras utilizadas para classificar os constituintes como segmentos ou subsegmentos, bem como, determinar quais os pontos, na estrutura textual, em que pode ocorrer uma segmentação. Na sequência, o módulo 2 recebe e reconhece os resultados apresentados no módulo 1, organizando os segmentos e subsegmentos de forma hierárquica em árvores DTS's, em conformidade com a aplicação das regras propostas para essa tarefa; a seguir, o módulo 3 acolhe, do módulo anterior, a organização arbórea com os segmentos e os subsegmentos organizados hierarquicamente, a partir desse resultado com os segmentos e os subsegmentos dispostos em DTS's mais as informações advindas do conjunto de regras, desenvolvido

especificamente para identificar as relações entre os segmentos, inicia-se a tarefa proposta para módulo 3, isto é, atribuir as relações retóricas entre os segmentos e os subsegmentos.

Concluída a revisão das etapas dos três módulos anteriores, chega-se à constituição do módulo 4, o qual reconhece, em toda a análise, a classificação, a organização segmentos – subsegmentos e as relações entre eles. Face às características observadas, a metodologia propõe que sejam a engendradas as informações dos níveis anteriores, relacionado-as com as regras propostas para a identificação das macroproposições localizadas, presentes em toda a superfície do texto. O objetivo desta manipulação de dados obtidos na execução dos módulos 1, 2 e 3 é prover o módulo 4 com dados a serem incorporados no conjunto com as regras que se destinam à composição e à sistematização da etapa destinada à geração automática da estrutura que representa conceitualmente o texto analisado, ou seja, a macroestrutura/macroproposição.

Dessa forma, o módulo 4 manipula os segmentos e subsegmentos devidamente identificados e dispostos em árvores, em que cada um dos constituintes desempenha um papel específico em relação ao tema do texto. A organização dos segmentos e dos subsegmentos é realizada no nível microestrutural, a partir de condições previamente determinadas, em cumprimento a um objetivo específico que o autor quer que se realize. Neste sentido, a análise realizada em nossa metodologia coaduna com a definição de Kintsch [18], que define a microestrutura como sendo uma base representativa de texto da ordem abstrata, composta por *proposições*, as quais são representadas por estruturas formais organizadas entre si de modo hierárquico, identificados em nossa investigação como segmentos e subsegmentos. A microestrutura é, desta forma, construída a partir de conceitos representados na superfície textual por palavras ou grupo de palavras, por frases inteiras, ou pelos constituintes textuais, segmentos e subsegmentos, os quais encontram-se concatenados por relações conceituais, isto é, as relações retóricas.

O objetivo deste módulo é apresentar uma estrutura representativa do tema de um texto, para tal, consideram-se os dados e as características identificadas no nível da microestrutura, a fim de se chegar ao nível da macroestrutura/macroproposição, isto é, o texto na sua totalidade significativa. Apropriando-se desse reconhecimento micro-macroestrutural, pretende-se gerar automaticamente, a partir da execução das etapas 1, 2 e 3 do sistema AuTema-Dis, uma estrutura que apresenta o tema do texto avaliado, pretende-se gerar a macroestrutura/macroproposição.

Segundo os autores Kintsch e Dijk [18], Dijk [12], a macroestrutura corresponde a um nível global de descrição que ultrapassa a estrutura semântica linear do discurso, isto é, a microestrutura, ainda que a sua significação dependa das proposições explicitadas na base do

texto e da sua referência, bem como, das relações que estas estabelecem entre si. No âmbito desta investigação, a caracterização e a definição apresentadas pelos autores correspondem adequadamente à proposta metodólogica desenvolvida para a execução do módulo 4, no articulação dos níveis micro e macroestruturais.

Na proposta dos autores, a macroestrutura global, isto é, estrutura completa, representativa da temática discursiva, é constituída por macroproposições localizadas, que se encontram diretamente expressas na superfície textual ou podem ser construídas a partir destas proposições da base do texto, reorganizadas e explicitadas através da aplicação de macrorregras, cuja função consiste em transformar a informação semântica e a unidade que expressa a macroestrutura, também identificada como *macroproposição textual.* A aplicação das macrorregras reduz e abstrai o conteúdo proposicional das sequências textuais e, ao mesmo tempo, organizam o seu conteúdo em termos de hierarquização, tais regras orientam a produção das macroestruturas/macroproposições nos textos, conforme apresentamos na figura 4.5.

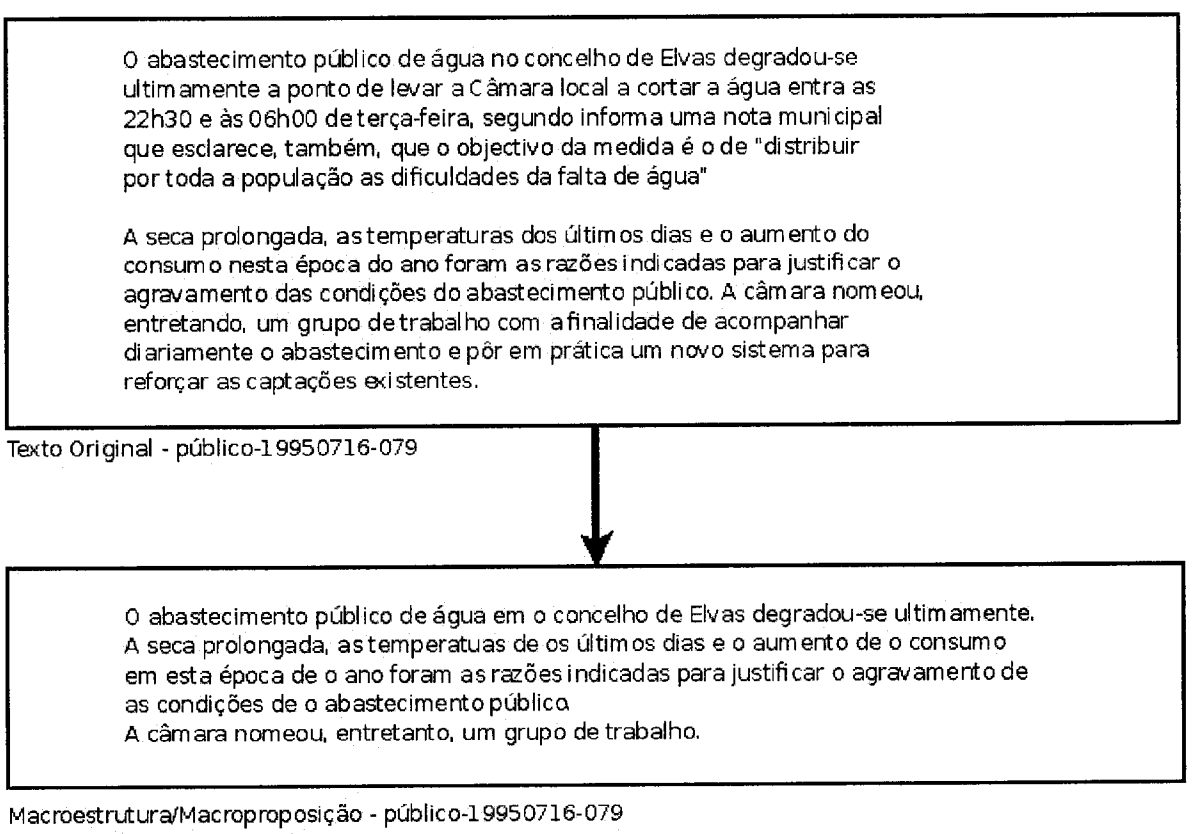

Figura 4.5: A figura representa a macroestrutura/macroproposição de um texto processado pelo AuTema-Dis –publico19950716-079.

4.2.4.1 As Macrorregras e os Níveis de Profundidade Textual dos Constituintes

Na metodologia proposta consideramos as macrorregras e os níveis de profundidade em que se encontram os constituintes textuais como elementos que auxiliam a identificação da macroestrutura/macroproposição em um texto. Dijk [10] apresenta essas macrorregras como operações que selecionam, reduzem, generalizam e (re-)constróem proposições em outras proposições menores, mais gerais ou mais particulares. As macrorregras identificadas por Dijk [10] são: *generalização* – que permite a substituição de sequências de uma ou mais proposições por uma proposição geral que as englobe; *apagamento/supressão* – que suprime as proposições não relevantes para a compreensão do texto; *integração/construção* – que permite integração de informações uma proposição de ordem inferior por uma unidade ou sequência mais ampla. Elas, segundo o autor, são regras de interpretação de sentenças e de pares de sentenças como proposições (globais), que caracterizam o significado de uma sequência de ações realizadas.

A aplicação das macrorregras auxiliam na supressão da informação proposicional de relevância exclusivamente local que não seja necessária para a compreensão do resto do discurso. Neste sentido corroboramos com Dijk e relacionamos a proposta das macrorregras na elaboração das regras utilizadas pelo sistema AuTema-Dis, que selecionam apenas as estruturas imprescindíveis à composição da macroestrutura/macroproposição global, no sentido de eliminar as informações que não sejam importantes à constituição temático-conceitual do discurso.

Ressaltamos que as macrorregras de Dijk não foram empregadas na sua essência em nossa investigação. Entretanto, nos forneceram orientações da ordem linguística-conceitual, que ajudaram a caracterizar as regras utilizadas para selecionar os constituintes textuais que não apresentam conteúdo semântico relacionado diretamente ao tema tratado no texto necessário à produção automática da macroestrutura/macroproposição. Para podermos aproveitar a caracterização proveniente da aplicação das macrorregras, foi necessário avaliamos o emprego das três macrorregras de Dijk em diferentes estudos, só assim foi possível validar a sua aplicablidade, no âmbito desta investigação. Especificamente no caso da construção do módulo 4, as macrorregras estão relacionadas aos níveis de profundidade em que se encontram os constituintes textuais no interior das estruturas que compõem os textos.

Definido o ponto de intersecção entre as macrorregras e nível de profundidade dos constituintes, definiu-se a edificação do módulo 4 que conta com a realização de duas tarefas: uma relacionada à identificação dos constituintes, a partir da aplicação das regras propostas no módulo 2; e outra responsável por avaliar a importância dos constituintes selecionados em relação a sua participação na composição conceitual de toda a estrutura. Para tal, foi

necessário considerar informação sobre o nível em que se encontram os constituintes, pois os níveis orientam a seleção e a categorização em relação ao papel de cada um dos constituintes na composição do tema, bem como, especifica qual a macrorregra a ser utilizada para a exclusão do constituinte.

Conforme evidenciamos nesta investigação, os níveis de profundidade relacionam-se às macrorregras no sentido que orientam a posição ocupada pelo subsegmento na estrutura, e qual macrorregra que poderá ser aplicada para organizar o conteúdo. Observou-se, pelas análises realizadas nos *corpora* que quanto mais interno estiver um subsegmento em uma estrutura, menor será o seu comprometimento com o tema global, isto é, quanto mais afastado estiver o subsegmento da estrutura que ocupa a posição do *nó principal*, ou nó de primeiro nível, maior será a probabilidade deste constituinte ser considerado descartável em relação à composição do tema. Neste sentido, alguns subsegmentos tornam-se candidatos a não participarem da composição da macroproposição/macroestrutura, sendo eliminados pela aplicação da macrorregra de apagamento ou supressão.

Considerando os resultados da análise realizada nos *corpora* e em conformidade com a possibilidade de implementação deste módulo em sistema computacional, optamos pela utilização de apenas uma das três macrorregras, isto é, utilizamos a macrorregra *apagamento ou supressão*. A opção pela aplicação da macrorregra de apagamento deve-se às restrições de implementação deste módulo em sistema computacional. As outras macrorregras são importantes, no entanto, no nível em que se encontra esse estudo e devido à complexidade que exige a implementação das outras duas macrorregras, optamos por trabalhar apenas com a que fosse imediatamente passível de implementação.

4.2.4.2 A Composição da Macroestrutura/Macroproposição

Conforme mencionamos, a seleção dos constituintes que participam da composição da macroestrutura/macroproposição textual é feita no módulo 4, a partir da realização de duas tarefas: identificação dos constituintes; eleição dos constituintes relacionados diretamente à temática do discurso. O processo de seleção dos segmentos e subsegmentos a participarem da composição da estrutura macropropositiva foi elaborado a partir das regras de seleção propostas não módulo 2, acrescidas pela informação sobre a posição/profundidade em que se encontra o constituinte e pelo princípio proposto por Dijk [10], para a aplicação das macrorregras, que suprime a informação proposicional de relevância exclusivamente local que não seja necessária para a compreensão do resto do discurso.

A macroestrutura/macroproposição implica coerência global, que confere ao texto a sua unidade, assegurando as ligações entre as várias partes que o constituem. Estas ligações entre os constituintes dos textos são consideradas na organização de macroestrutura/macroproposição dos textos, as ligações são contempladas nesta metodologia no módulo 3, identificadas através das relações retóricas. Acreditamos que as relações retóricas são capazes de auxiliar na explicação de como a coerência pode ser estabelecida no texto, bem como, contribuem para composição do sentido do texto, pois perpassam por toda a superfície textual estabelecendo relações micro e macroestruturais. No nível microestrutural evidenciam-se as relações sentenciais e intersentenciais organizadas a partir da inter-relação dos níveis morfo-sintático-semântico das estruturas localizadas. No nível macroestrutural, observam-se as relações entre os blocos constitutivos de texto e o seu resultado na composição do significado global de toda a estrutura.

Em termos de macroestrutura, observam-se as estruturas semânticas de um nível mais elevado que são derivadas das sequências proposicionais (macroproposições locais) do texto e das relações retóricas que se manifestam entre elas. As macroestruturas definem intuitivamente a noção de significado global, tema ou tópico de um texto ou fragmento de um texto. Desta forma, essas *macroestruturas* não são definidas unicamente em termos de significado sentencial ou sequências individuais, as macroestruturas organizam os significados ao longo do texto e o significado do texto como um todo. É neste sentido que a metodologia apresentada para o módulo 4 propõe que sejam selecionados os constituintes mais significativos para compor a *macroestrutura/macroproposição* do texto analisado.

Conforme Dijk [8], uma análise discursiva completa requer, além da análise relativa das sentenças do texto, uma descrição explícita das estruturas das sequências das sentenças. Ao descrevermos estas sequências, por exemplo quando identificamos as relações retóricas entre os segmentos de um texto, estamos identificado os processos de conexão entre as proposições do discurso no nível conceitual, e os processos que condicionam essas relações. Desta forma, ao elaborar a metodologia proposta a compor o módulo 4, procurou-se contemplar a articulação desses dois níveis de análise: um que recorre ao texto enquanto representação do discurso e o outro ao próprio discurso que se manifesta concretamente a partir do texto.

Em um sentido pragmático o módulo 4 prevê a seleção dos constituintes que ocupam os *nós* de primeiro nível e alguns do nível 2, de acordo com as regras de eleição dos segmentos e subsegmentos. Selecionados os constituintes, organiza-se a macroestrutura/macroproposição do texto analisado. A constituição da metodologia do módulo 4 foi elaborada, como os módulos 1, 2 e 3, para ser implementada e testada em sistema computacional. Os textos são avaliados enquanto estruturas resultantes das relações morfo-sintático-semânticas, a partir de regras especificamente elaboradas e apresentada na descrição de cada um dos módulos.

4.3 Resumo do Capítulo

Neste capítulo, apresentamos a metodologia desenvolvida para realizar análise discursiva completa. A proposta metodológica foi construída a partir da edificação de quatro módulos independentes que compartilham resultados, para a constituição e validação de um processo unificado. Desta forma, apresentamos e descrevemos a nossa proposta metodológica considerando cada um dos quatro módulos que a constitui, são eles:

1. Módulo 1 - Identificação e Segmentação dos Constituintes Textuais
2. Módulo 2 - Organização Arbórea -DTS's
3. Módulo 3 - Identificação das Relações Retóricas em DTS´s
4. Módulo 4 - Representação Estrutural da Macroproposição Textual

4.3.1 Metodologia Modular - Implementação AuTema-Dis

A metodologia proposta e apresentada nas seções anteriores foi prevista para ser demonstrada e avaliada a partir de uma componente sistemática. Assim, no próximo capítulo desta tese, apresentaremos a implementação de cada um dos módulos previstos na metodologia para análise textual em sistema computacional, nomeadamente, AuTema-Dis. A implementação realizada destina-se a avaliar a metodologia proposta e a sua execução automática, a fim de poder oferecer suporte a novas análises, bem como, ser possível ampliar a proposta metodológica em questão.

Capítulo 5

AUTEMA-DIS: Avaliação e Aplicações

O processo de avaliação de cada um dos quatro módulos para análise textual propostos na metodologia é realizado sistematicamente. Neste sentido, respeitamos os processos que envolvem a implementação em sistema computacional, concernentes à edificação do AuTema-Dis, bem como, os resultados advindos da sua execução. A avaliação de cada um dos módulos previstos é realizada de duas formas, isto é, manual e automática, e os resultados podem ser evidenciados, conforme demonstramos neste capítulo, em tabelas com dados estatísticos.

Outrossim, ressaltamos que o processo de avaliação realizado no âmbito desta tese foi concebido de forma especial, permitindo-nos dar cobertura a dois processos distintos, isto é:

- elaboração da metodologia proposta para análise discursiva automática;
- constituição e execução do sistema computacional, AuTema-Dis, estruturado a partir da metodologia proposta.

Cabe-nos ainda ressaltar que, apesar de realizarmos dois processos distintos de avaliação, para fins de organização descritiva, os mesmos serão apresentados inter-relacionados, podendo não haver em alguns pontos do texto uma distinção nitidamente formal entre a apresentação de um e outro processo avaliativo.

5.1 Avaliação dos Módulos

Conforme apresentamos no capítulo 4, propusemos uma metodologia do tipo modular em que foram previstos quatro módulos distintos para realizar análise textual, idealizando-se a

possibilidade da sua implementação em sistema computacional. A execução computacional é a forma pragmática empregada para avaliarmos e validarmos a metodologia desenvolvida, especificamente, para análise automática discursiva.

5.1.1 Medidas de Avaliação

A metodologia proposta nesta tese foi avaliada considerando-se as seguintes medidas: precisão *–precision*, cobertura *–recall* e *F-measure*, as quais apresentamos a definição na sequência.

A precisão *(P)* corresponde à exatidão do sistema em realizar uma tarefa específica, por exemplo, encontrar os segmentos e subsegmentos existentes em um texto. Para identificarmos a precisão do sistema, utilizamos uma fórmula em que o número de segmentos e subsegmentos corretos encontrados automaticamente pelo sistema é dividido pelo número de segmentos e subsegmentos identificados automaticamente. A medida da Precisão é também utilizada para identificar: a segmentação dos constituintes textuais; organização arbórea dos constituintes textuais em DTS's; e a atribuição das relações retóricas entre os constituintes.

Especificamente ao caso da avaliação da macroestrutura/macroproposição, não foi possível aplicar a medida de precisão, por se tratar de uma produção de caráter subjetivo. Assim, a avaliação da execução do sistema na realização do módulo 4, que concerne à apresentação automática da estrutura temática do texto, foi realizada em termos de coerência/incoerência e completude/incompletude, conforme descrevemos na sequência deste capítulo, especificamente na *Avaliação do módulo 4*.

A cobertura *(C)* corresponde à abrangência do sistema em realizar uma determinada tarefa (identificação dos segmentos, segmentação dos constituintes textuais, organização arbórea, atribuição das relações retóricas). Para identificarmos a cobertura apresentada pelo sistema, por exemplo, no que se refere à identificação automática dos segmentos e subsegmentos, utilizamos uma fórmula dividindo o número de segmentos e subsegmentos corretos encontrados automaticamente pelo número de segmentos identificados manualmente. Assim, a partir da aplicação desta medida, podemos identificar o quão completo é a execução do sistema a identificar corretamente os segmentos e os subsegmentos.

No caso da medida F–measure *(F)*, trata-se de uma média harmônica ponderada entre precisão (exatidão do sistema em realizar uma tarefa) e a cobertura (abrangência do sistema em realizar uma determinada tarefa). A F-measure apresenta o desempenho do sistema em, por

exemplo, identificar os segmentos e subsegmentos e, a partir dos resultados desta identificação, verificar quantos destes elementos foram corretamente apresentados pelo sistema.

Para medir o desempenho do sistema, utilizamos também as medidas *macromédia* e *micromédia*, conforme a proposta de Joachims [16]. Tais medidas são comumente usadas para computar o desempenho médio do sistema sobre múltiplos conjuntos de dados, em que os resultados de n tarefas são mensuradas para indicar um valor de desempenho único. A *Macromédia* corresponde à forma padrão de contabilizar uma média aritmética. O desempenho (precisão e cobertura) são verificados separadamente para cada um dos n testes. A média apresenta de forma aritmética o desempenho sobre todos os testes. A *Micromédia* apresenta o resultado da média para cada um dos exemplos presentes na tabela. Para cada célula da tabela, é feita a média aritmética e o desempenho é verificado a partir destas médias. Enquanto a macromédia atribui um peso igual para cada teste, micromédia considera cada exemplo do teste.

Na sequência desta seção, apresentaremos passo-a-passo a avaliação realizada para cada um dos módulos propostos na metodologia devidamente implementados e testados no sistema AuTema-Dis, conforme segue.

5.1.2 Avaliação do Módulo 1 – Identificação e Segmentação Automática dos Constituintes Textuais –

O módulo 1, conforme descrevemos, apresenta a metodologia para realizar automaticamente a identificação dos constituintes textuais e a sua posterior segmentação em unidades significativas, em conformidade com sua participação na composição temática do discurso. Assim, o módulo 1 é constituído pelo conjunto de regras de cunho morfo-sintático-semântico elaboradas a partir dos resultados da análise manual, realizada pelo analista e da análise automática, executada pelo analisador *Palavras* nos textos que fazem parte dos *corpora*.

A avaliação da metodologia proposta para o módulo 1 é realizada considerando-se: as regras elaboradas para identificação e segmentação dos constituintes textuais, implementadas em sistema computacional e a sua posterior execução e validação dos resultados obtidos. A implementação das regras compõem a primeira etapa do sistema que realiza dois processos diferenciados, mas complementares: identificação dos segmentos e subsegmentos e a sua posterior segmentação em unidades mínimas de significação.

A partir da execução em sistema computacional do módulo 1, realizamos a avaliação a fim de verificar se o sistema reconhece os segmentos e subsegmentos e, considerando os resultados desta identificação como uma pré-etapa para a seguinte, se ele é capaz de segmentar um texto de acordo com a classificação dos constituintes. Assim, para avaliarmos essas duas atividades realizadas pelo sistema, recorremos à análise manual e automática de todos os 50 textos constituintes dos *corpora*.

Realizamos inicialmente a identificação manual dos segmentos e subsegmentos nos 10 textos em *PE*, os quais fazem parte do conjunto *aprendizado/treino*. Na sequência, as mesmas regras de identificação foram implementadas e os resultados obtidos manual e automaticament podem ser observados nas tabelas 5.1 e 5.2. Frente aos resultados obtidos a partir da análise realizada, desenvolvemos uma atividade semelhante para os outros dois *corpora* constituídos por 20 textos em *PB* e 20 textos em *PE*, os quais perfazem o conjunto *avaliação/teste*. Na continuidade do processo, utilizamos a mesma sistemática de análise para a totalidade dos *corpora*. Assim, o módulo 1, devidamente implementado no sistema AuTema-Dis, realizou automaticamente a identificação dos segmentos e subsegmentos,na totalidade dos corpora, o que constitui a 1ª etapa do AuTema-Dis.

Os resultados do processamento da 1ª etapa do processamento do sistema AuTema-Dis foram apresentados nas tabelas 5.3, 5.4, 5.5 e 5.6 em que é possível verificar o número de segmentos e subsegmentos identificados manual e automaticamente, bem como, a correção na identificação dos constituintes pelo sistema AuTema-Dis. Para avaliarmos a correção, utilizamos a identificação manual como base para determinarmos se o sistema correspondeu adequadamente.

No que se refere à avaliação da metodologia proposta para identificação e segmentação textual, realizamos uma análise contrastiva entre a identificação realizada manualmente e a executada pelo analisador *AuTema-Dis*. O analisador desenvolvido tem como base as regras de identificação e segmentação, sendo constituído parcialmente por características do analisador *Palavras*, descritas no 4° capítulo.

Nas tabelas 5.1 e 5.2 é possível observar a avaliação do primeiro conjunto de textos, isto é, o conjunto *aprendizado/treino*, constituído por 10 textos do Jornal Público. No que diz respeito à correção da metodologia proposta para identificação automática dos segmentos e subsegmentos, verifica-se que a automatização desta primeira etapa apresenta resultados muito satisfatórios, principalmente na identificação dos segmentos. É possível observar em termos percentuais que o sistema apresenta uma correção na identificação dos segmentos de 82% e de 70% na identificação dos subsegmentos.

Corpus Aprendizado - Jornal Público - 1994 - 1995						
Conjunto de Textos	Segmentos Identificados					
	Manual	Automático	Correção		Erro	
					Analisador	Regra
			%	n	n	n
Nº 19940101-007	4	4	100%	4	0	0
Nº 19941012-035	2	2	50%	1	1	0
Nº 19941214-076	6	6	83%	5	1	0
Nº 19950519-057	3	3	33%	1	1	1
Nº 19950725-025	2	2	100%	2	0	0
Nº 19950726-079	3	3	100%	3	0	0
Nº 19950916-121	5	5	80%	4	0	1
Nº 19950916-157	7	7	100%	7	0	0
Nº 19950917-041	5	5	80%	4	1	0
Nº 19950814-011	5	5	80%	4	1	0
Média - Totais	4.2	4.2	81%	4	0.5	0.2

Tabela 5.1: A tabela apresenta os resultados manual e automático obtidos na identificação e classificação dos segmentos nos textos do conjunto aprendizado.

Corpus Aprendizado - Jornal Público - 1994 - 1995						
Conjunto de Textos	Subsegmentos Identificados					
	Manual	Automático	Correção		Erro	
					Analisador	Regra
			%	n	n	n
Nº 19940101-007	13	13	100%	13	0	0
Nº 19941012-035	5	6	83%	5	1	0
Nº 19941214-076	12	8	100%	8	0	0
Nº 19950519-057	5	6	50%	3	2	1
Nº 19950725-025	7	5	60%	3	2	0
Nº 19950726-079	5	6	83%	5	0	1
Nº 19950916-121	8	6	33%	2	3	1
Nº 19950916-157	9	7	71%	5	0	2
Nº 19950917-041	5	4	75%	3	1	0
Nº 19950814-011	8	2	0%	0	3	3
Média - Totais	7.7	6.3	66%	4.7	1.2	0.8

Tabela 5.2: A tabela apresenta os resultados manual e automático obtidos na identificação e classificação dos subsegmentos nos textos do conjunto aprendizado.

Percebe-se claramente que existe um declínio no percentual de correção do processo automático em relação ao processo manual na identificação dos segmentos e dos subsegmentos. Na análise realizada, justificamos esse declínio no percentual da correção devido à complexidade na organização estrutural constituída por subsegmentos, bem como, a evidentes falhas, cuja origem está na identificação realizada pelo analisador *Palavras*, estágio que faz parte da 1ª etapa do processamento AuTema-Dis.

No tocante aos demais textos do Jornal Público e Folha de São Paulo, os quais fazem parte dos conjuntos *avaliação/teste*, o procedimento de análise foi realizado de forma semelhante ao processo descrito para o conjunto *aprendizado/treino*. Assim sendo, as regras propostas na metodologia para o reconhecimento automático dos segmentos e subsegmentos foram avaliadas a partir da comparação entre o resultado da atividade realizada manualmente e o resultado da automatização no sistema AuTema-Dis, devidamente implementado com tais regras para executar a mesma atividade.

Nas tabelas 5.3, 5.4, 5.5 e 5.6 é possível verificar o contraste entre os resultados da identificação dos segmentos e subsegmentos realizados manualmente pelo analista e os resultados da implementação e execução das regras propostas para este fim, realizadas no sistema automático AuTema-Dis.

Corpus - Jornal Público - 1994 - 1995						
Conjunto de Textos	Segmentos Identificados					
	Manual	Automático	Correção		Erro	
					Analisador	Regra
			n	%	n	n
Nº 19940504-070	3	3	3	100%	0	0
Nº 19940505-024	4	4	2	50%	1	1
Nº 19940505-071	4	4	2	50%	1	1
Nº 19941911-083	7	7	5	71%	1	1
Nº 19941012-011	5	5	5	100%	0	0
Nº 19941025-045	6	6	4	67%	1	1
Nº 19950416-032	4	4	2	50%	2	0
Nº 19950795-167	6	6	3	50%	3	0
Nº 19950912-022	4	4	3	75%	0	1
Nº 19950924-121	11	11	8	73%	1	2
Nº 19950422-141	4	4	3	75%	0	1
Nº 19950423-011	5	4	3	75%	1	0
Nº 19950629-083	5	5	4	80%	1	0
Nº 19950629-119	6	7	4	57%	2	1
Nº 19951011-139	2	2	2	100%	0	0
Nº 19951011-150	8	8	8	100%	0	0
Nº 19951114-163	2	2	1	50%	0	1
Nº 19951114-169	6	6	3	50%	1	2
Nº 19951220-045	4	4	1	25%	2	1
Nº 19951229-044	4	4	4	100%	0	0
Média Totais	5	5	3.5	70%	0.85	0.65

Tabela 5.3: A tabela apresenta os resultados manual e automático obtidos na identificação e classificação dos segmentos nos textos do conjunto avaliação/teste do Jornal Público.

Corpus - Jornal Público - 1994 - 1995						
Conjunto de Textos	Subsegmentos Identificados					
	Manual	Automático	Correção		Erro	
					Analisador	Regra
			n	%	n	n
Nº 19940504-070	6	6	4	67%	1	1
Nº 19940505-024	7	5	3	60%	1	1
Nº 19940505-071	8	9	8	89%	0	1
Nº 19941911-083	9	13	9	69%	3	1
Nº 19941012-011	11	9	7	78%	2	0
Nº 19941025-045	15	18	15	83%	2	1
Nº 19950416-032	9	6	3	50%	3	0
Nº 19950795-167	16	5	5	100%	0	0
Nº 19950912-022	4	9	2	22%	5	2
Nº 19950924-121	11	13	10	77%	1	2
Nº 19950422-141	9	8	3	38%	4	1
Nº 19950423-011	6	7	4	57%	3	0
Nº 19950629-083	8	10	8	80%	1	1
Nº 19950629-119	13	16	11	69%	4	1
Nº 19951011-139	10	6	0	0%	5	1
Nº 19951011-150	7	4	1	25%	2	1
Nº 19951114-163	5	5	5	100%	0	0
Nº 19951114-169	9	5	1	20%	2	2
Nº 19951220-045	9	6	3	50%	2	1
Nº 19951229-044	12	11	11	100%	1	0
Média - Totais	9.2	8.55	5.65	66%	2.1	0.9

Tabela 5.4: A tabela apresenta os resultados manual e automático obtidos na identificação e classificação dos subsegmentos nos textos do conjunto avaliação/teste do Jornal Público.

Conforme apresentamos, a avaliação das regras desenvolvidas para identificar automaticamente os constituintes textuais e segmentá-los foi realizada de duas formas:

1. uma manual, realizada pelo analista, o qual avalia as regras propostas manualmente em cada um dos textos dos corpora;

2. uma automática, em que as regras já se encontram implementadas em um sistema computacional (AuTema-Dis), desenvolvido para o teste, em que os textos dos corpora são executados.

No que diz respeito à avaliação dos resultados, no processo de implementação das regras propostas para identificação dos constituintes na totalidade do conjunto *avaliação/treino*, verificamos que o sistema responde de forma satisfatória à correção dos segmentos e subsegmentos encontrado, apesar de apresentar uma diferença nos resultados obtidos para os textos em *PB* e para os em *PE*.

Corpus - Folha de São Paulo - 1994 - 1995						
Conjunto de Textos	Segmentos Identificados					
	Manual	Automático	Correção		Erro	
					Analisador	Regra
			n	%	n	n
Nº FSP950101-011	5	4	4	100%	0	0
Nº FSP950101-032	4	4	3	75%	1	0
Nº FSP950101-054	6	5	4	80%	1	0
Nº FSP950101-084	13	13	6	46%	6	1
Nº FSP950111-014	5	5	5	100%	0	0
Nº FSP950111-026	6	7	4	57%	2	1
Nº FSP950111-034	9	9	9	100%	0	0
Nº FSP950111-036	2	2	2	100%	0	0
Nº FSP950117-048	9	9	7	78%	1	1
Nº FSP950117-074	8	6	4	67%	1	1
Nº FSP940101-132	4	4	2	50%	2	0
Nº FSP940101-124	10	9	6	67%	3	0
Nº FSP940101-107	12	12	9	75%	2	1
Nº FSP940101-102	12	11	9	82%	2	0
Nº FSP940101-095	22	20	20	100%	0	0
Nº FSP940101-092	8	8	5	63%	1	2
Nº FSP940101-085	7	7	5	71%	2	0
Nº FSP940101-079	4	4	4	100%	0	0
Nº FSP940101-074	10	10	9	90%	1	0
Nº FSP940101-066	9	9	8	89%	1	0
Média - Totais	8.25	7.9	6.25	79%	1.3	0.35

Tabela 5.5: A tabela apresenta os resultados manual e automático obtidos na identificação e classificação dos segmentos nos textos do conjunto avaliação/teste da Folha de São Paulo.

No tocante à identificação específica dos *segmentos*, a execução das regras pelo sistema AuTema-Dis apresentou resultados mais precisos para textos em *PB*, em comparação com os textos em *PE*. A correção do sistema para o *PB* foi de quase o dobro de segmentos corretos identificados se compararmos com os resultados para o *PE*. A média de segmentos identificados corretamente pelo sistema para o *PB* foi de 6.25, enquanto que para os textos em *PE* a média ficou em torno de 3.5, apresentando uma diferença de 2.75 na correção. Acreditamos que o sistema tenha um desempenho melhor para os textos em *PB* devido a dois fatores:

- a primeira etapa do sistema é constituída por regras, destinada à segmentação dos constituintes textuais, as quais foram elaboradas a partir da realização da análise do *Palavras*. A composição do analisador *Palavras* foi estruturada para o Português Brasileiro, o que nos leva a considerar esse melhor desempenho devido as bases linguís-

Corpus - Folha de São Paulo - 1994 - 1995						
Conjunto de Textos	Subsegmentos Identificados					
	Manual	Automático	Correção		Erro	
					Analisador	Regra
			n	%	n	n
Nº FSP950101-011	16	13	11	85%	2	0
Nº FSP950101-032	10	11	10	91%	1	0
Nº FSP950101-054	3	3	3	100%	0	0
Nº FSP950101-084	15	12	7	58%	5	0
Nº FSP950111-014	7	5	1	20%	2	2
Nº FSP950111-026	11	12	7	58%	3	2
Nº FSP950111-034	9	10	7	70%	2	1
Nº FSP950111-036	6	3	3	100%	2	1
Nº FSP950117-048	10	18	8	44%	8	2
Nº FSP950117-074	13	9	3	33%	4	2
Nº FSP940101-132	6	6	4	67%	2	0
Nº FSP940101-124	15	8	1	13%	5	2
Nº FSP940101-107	7	6	5	83%	1	0
Nº FSP940101-102	10	10	9	90%	0	1
Nº FSP940101-095	19	21	15	71%	5	1
Nº FSP940101-092	11	11	11	100%	0	0
Nº FSP940101-085	12	10	7	70%	2	1
Nº FSP940101-079	15	17	15	88%	2	0
Nº FSP940101-074	10	10	4	40%	5	1
Nº FSP940101-066	10	11	8	73%	3	0
Média - Totais	10.75	10.3	6.95	67%	2.7	0.8

Tabela 5.6: A tabela apresenta os resultados manual e automático obtidos na identificação e classificação dos subsegmentos nos textos do conjunto avaliação/teste da Folha de São Paulo.

tico-gramaticais do analisador.

- as estruturas que compõem os textos em PB são menos complexas no que diz respeito a organização estrutural e sintática do que os textos escritos em PB.

A avaliação do sistema AuTema-Dis para a identificação dos *subsegmentos* no conjunto *avaliação/treino* apresentou as mesmas características observadas nos resultados da avaliação na identificação dos *segmentos*. Verificou-se que o sistema AuTema-Dis realiza de forma satisfatória e equilibrada a identificação dos constituintes nos textos em *PB*, perfazendo uma média de subsegmentos corretos em *PB* de 6.95, enquanto que a correção nos textos em *PE* perfaz 5.65. A média na identificação dos subsegmentos elevou-se no caso dos textos em *PE* aproximando-se da média de identificação para os textos em *PB*. Neste sentido, avaliamos positivamente essa aproximação das médias no que diz respeito ao correto reconhecimento dos subsegmentos pelos sistema.

Notadamente, o sistema AuTema-Dis executa de forma satisfatória e equilibrada a tarefa na identificação dos segmentos e subsegmentos nos textos. No entanto, percebe-se que apesar das regras para identificação dos segmentos terem sido elaboradas a partir de um conjunto de 10 textos escritos em *PE*, a melhor performance e precisão do sistema ocorre em relação aos textos escritos em *PB*. Acreditamos que essa diferença está pautada na organização das estruturas constituintes dos textos.

5.1.2.1 As regras de Segmentação

As regras para executar a segmentação dos textos em constituintes foram desenvolvidas na metodologia e são utilizadas na elaboração da 1ª etapa da análise realizada pelo AuTema-Dis. A implementação do conjunto das regras permite que os textos sejam segmentados em unidades mínimas significativas *EDU's*, isto é, entidades que representam papéis diferenciados na composição do tema do texto.

Nesta seção, apresentamos a avaliação das regras destinadas à segmentação automática dos constituintes textuais em duas tabelas 5.7 e 5.8. Nas tabelas priorizamos a identificação das ocorrências por regra proposta para segmentação. A tabela 5.7 apresenta o conjunto das regras de segmentação evidenciadas inicialmente a partir do conjunto *aprendizado/treino*, em *PE*. Salientamos que o conjunto inicial apresenta um número mais reduzido de regras, pois estas foram constituídas a partir das características evidenciadas em apenas 10 textos, os quais constituem o conjunto *aprendizado/treino*, conforme tabela 5.7.

Regras de Segmentação - Avaliação 10 Textos - PE		
Regras Iniciais	Nº de Ocorrências Manual	Nº Ocorrências Sistema
N<:fcl	5	5
Advl:pp	42	35
Advl:advp	7	7
Advl:fcl	1	2
Pred:pp	6	3
Advl:cu	1	1
App:prop	1	1
Co:conj-c ('mas')	3	2
Pred:np;	2	2
App:np	2	2
Advl:adv - atemp	6	1
Advl:np ou Advl:n	1	2

Tabela 5.7: A tabela apresenta as ocorrências manual e automática das regras de segmentação textual no conjunto aprendizado, constituído por dez textos em Português Europeu.

Na análise da tabela 5.7, identifica-se, no entanto, que existe um equilíbrio entre a segmentação manual dos constituintes textuais com a segmentação automática, realizada pelo AuTema-Dis. O sistema implementado responde adequadamente ao reconhecimento dos constituintes nos 10 textos iniciais.

A segmentação realizada manualmente na totalidade dos corpora, isto é, nos 50 textos que os constituem, revelou ser necessária a inclusão de mais regras para dar conta do processo de segmentação, conforme apresentamos na tabela 5.8. Ao compararmos ambas tabelas com a ocorrência das regras é possível identificar que foram incluídas mais 3 regras de segmentação. A necessidade de inclusão de mais regras foi condicionada pelo aumento no número de textos analisados, bem como, à diversidade na constituição das estruturas dos textos dos corpora analisados, isto é, em *PB* e *PE*.

Regras de Segmentação - Avaliação 50 Textos		
Regras Finais	Nº de Ocorrências Manual	Nº Ocorrências Sistema
N<:fcl	38	40
Advl:pp	239	205
Advl:advp	15	16
Advl:fcl	30	34
Pred:pp	19	16
Advl:cu	7	7
App:prop	16	14
Advl:acl	8	8
Sta:icl	10	13
Co:conj-c ('mas')	14	12
Pred:np;	13	7
App:np	8	10
Advl:adv - atemp	52	47
Advl:adv - aloc	1	3
Advl:np ou Advl:n	6	8

Tabela 5.8: A tabela apresenta as ocorrências manual e automática com regras de segmentação textual, identificadas na totalidade dos corpora.

Para avaliarmos adequadamente o desempenho do sistema AuTema-Dis no processo de segmentação textual, efetuamos a comparação dos resultados obtidos entre a segmentação manual e a segmentação automática, utilizando-se como referência a identificação manual realizada pelo analista. Desta forma, foi possível verificar a cobertura e a precisão alcançadas pelo sistema AuTema-Dis, considerando-se, para tal, o conjunto de regras propostas na metodologia, conforme apresentamos na tabela 5.8.

Na sequência do processo avaliativo é possível identificar as tabelas 5.9 e 5.10 representam

estatisticamente os percentuais para o desempenho do sistema AuTema-Dis no tocante à identificação e segmentação de textos.

Segmentos - Totalidade dos Corpora			
	Precisão	Cobertura	F-Measure
Macromédia	0.76	0.75	0.75
Micromédia	0.77	0.75	0.76

Tabela 5.9: A tabela apresenta os resultados estatísticos do sistema Autema-Dis na identificação e classificação automática dos segmentos nos textos dos corpora.

Subsegmentos - Totalidade dos Corpora			
	Precisão	Cobertura	F-Measure
Macromédia	0.65	0.63	0.63
Micromédia	0.73	0.67	0.70

Tabela 5.10: A tabela apresenta os resultados estatísticos do sistema Autema-Dis na identificação e classificação automática dos subsegmentos nos textos dos corpora.

Observa-se, a partir dos resultados demonstrados, que as regras propostas para identificação e segmentação dos constituintes textuais, devidamente implementados em sistema automático para análise textual, respondem satisfatoriamente em termos de desempenho, precisão e cobertura. No caso da identificação dos *segmentos* a precisão do sistema apresenta uma média próxima a 0.80, verificando-se os mesmos valores para a cobertura e a f-measure. No que diz respeito ao desempenho do AuTema-Dis noprocesso de identificação dos *subsegmentos*, verificou-se um decréscimo na média próxima a 0.70. O decréscimo pode ser justificado devido à complexidade em se analisar automaticamente os níveis mais profundos das estruturas.

A fim de nos certificarmos da viabilidade das regras desenvolvidas para realizar a segmentação dos textos, optamos por realizar uma análise considerando a execução de cada uma das regras, conforme pode ser evidenciado na tabela 5.11. A identificação por regras permitiu-nos, por exemplo, excluir as regras que apresentavam apenas uma ocorrência na totalidade dos corpora. Esta ação foi importante no sentido de contornarmos alguns problemas, tais como o nível de granularidade na segmentação dos textos, bem como, na identificação e delimitação das fronteiras/limites dos segmentos, processo esse que é responsável por alguns casos de ambigüidade na etapa da identificação das relações retóricas.

A comparação entre as análises demonstra que ambas identificações, manual e automática, apresentam-se equilibradas, o que pode ser evidenciado pelos resultados demonstrados na tabela 5.11. O percentual de correção do sistema apresenta um resultado satisfatório e harmônico na correta atribuição/identificação das regras de subsegmentação. Ressaltamos

Regras de Segmentação Segmentação	Avaliação das Regras							
	10 Textos		Correção		50 Textos		Correção	
	Manual	Sistema	%	n°	Manual	Sistema	%	n°
N<:fcl	5	5	100%	5	38	40	72%	29
Advl:pp	42	35	69%	24	239	205	72%	148
Advl:advp	7	7	71%	5	15	16	94%	15
Advl:fcl	1	2	50%	1	30	34	76%	26
Pred:pp	6	3	100%	3	19	16	81%	13
Advl:cu	1	1	100%	1	7	7	100%	7
App:prop	1	1	100%	1	16	14	86%	12
Advl:acl	0	0	0%	0	8	8	38%	3
Sta:icl	0	0	0%	0	10	13	62%	8
Co:conj-c ('mas')	3	2	100%	2	14	12	92%	11
Pred:np;	2	2	100%	2	13	7	0%	0
App:np	2	2	100%	2	8	10	80%	8
Advl:adv - atemp	6	1	100%	1	52	47	53%	25
Advl:adv - aloc	0	0	0%	0	1	3	0%	0
Advl:np ou Advl:n	1	2	50%	1	6	8	38%	3
Média	6.42	5.25	87%	4.00	31.73	29.33	73%	20.53

Tabela 5.11: A tabela apresenta o número de ocorrências por regras de segmentação na totalidade dos corpora, o contraste entre a identificação manual e a identificação automática e o percentual de correção do sistema.

que os percentuais encontram-se próximos a 87% para o conjunto *aprendizado/treino* e 73% para o conjunto *avaliação/teste*, o que identificamos como um bom resultado considerando-se o tipo de proposta em questão.

Os resultados estatísticos para os grupos *aprendizado* e *avaliação* no que se refere à aplicabilidade das regras, podem ser evidenciados nas tabelas 5.12 e 5.13. Percebe-se que a precisão, a cobertura e a f-measure em ambos conjuntos apresentam um equilíbrio nos resultados das médias em torno de 60 e 70%. Neste sentido, podemos concluir que os resultados da execução do sistema são compatíveis com os resultados da análise manual, justificando-se, desta forma, a regularidade do AuTema-Dis em realizar a tarefa proposta.

A tabela referida demonstra que a atribuição adequada das regras proporciona uma segmentação mais correta dos constituintes, o que por consequência pode favorecer a execução das 2ª e 3ª etapas do processamento do AuTema-Dis, nas quais os segmentos e subsegmentos são organizados automaticamente em DTS's e, recebem a classificação das relações retóricas.

Regras de Segmentação 10 Textos	Precisão	Cobertura	F-Mesure
N<:fcl	1.0	1.0	1.0
Advl:pp	0.7	0.6	0.6
Advl:advp	0.7	0.7	0.7
Advl:fcl	0.5	1.0	0.7
Pred:pp	1.0	0.5	0.7
Advl:cu	1.0	1.0	1.0
App:prop	1.0	1.0	1.0
Advl:acl	0.0	0.0	0.0
Sta:icl	0.0	0.0	0.0
Co:conj-c ('mas')	1.0	0.7	0.8
Pred:np;	1.0	1.0	1.0
App:np	1.0	1.0	1.0
Advl:adv - atemp	1.0	0.2	0.3
Advl:adv - aloc	0.0	0.0	0.0
Advl:np ou Advl:n	0.5	1.0	0.7
Macromédia	0.87	0.80	0.70
Micromédia	0.76	0.62	0.69

Tabela 5.12: A tabela apresenta os resultados estatísticos relativos à execução do sistema no conjunto *aprendizado* no processo de segmentação dos constituintes textuais.

Regras de Segmentação 50 Textos	Precisão	Cobertura	F-Mesure
N<:fcl	0.72	0.76	0.74
Advl:pp	0.72	0.62	0.67
Advl:advp	0.94	1.00	0.97
Advl:fcl	0.76	0.87	0.81
Pred:pp	0.81	0.68	0.74
Advl:cu	1.00	1.00	1.00
App:prop	0.86	0.75	0.80
Advl:acl	0.38	0.38	0.38
Sta:icl	0.62	0.80	0.70
Co:conj-c ('mas')	0.92	0.79	0.85
Pred:np;	0.00	0.00	0.00
App:np	0.80	1.00	0.89
Advl:adv - atemp	0.53	0.48	0.51
Advl:adv - aloc	0.00	0.00	0.00
Advl:np ou Advl:n	0.38	0.50	0.43
Macromédia	0.73	0.74	0.73
Micromédia	0.70	0.65	0.67

Tabela 5.13: A tabela apresenta os resultados estatísticos relativos à execução do sistema no conjunto *avaliação* no processo de segmentação dos constituintes textuais.

5.1.3 Avaliação do Módulo 2 – Organização Automática das DTS's – árvores de dependência dos segmentos –

O módulo 2 apresenta a metodologia desenvolvida para realizar computacionalmente a classificação e a organização dos constituintes textuais em árvores do tipo DTS's, previamente

identificados no módulo 1. A ordenação dos constituintes em árvores está condicionada ao papel que esses elementos desempenham em relação à composição do tema do texto, isto é, se são segmentos ou subsegmentos e ao nível de profundidade em que os constituintes se encontram nas estruturas.

A partir da implementação computacional da metodologia proposta no módulo 2, foi possível executar automaticamente a identificação e a organização dos segmentos e subsegmentos nas DTS's sem a interferência humana. Para estruturar os constituintes em DTS's, o sistema faz uso *a priori* dos resultados da execução computacional do módulo 1. Especificamente, na sequência do processamento, o módulo 2 recebe os resultados da etapa concluída no módulo 1, a qual identifica os constituintes textuais sem classificá-los, transformando esses resultados, até então não categorizados, em um novo conjunto de regras, as quais são utilizadas na geração automática da estrutura arbórea.

A edificação das árvores é feita a partir das regras propostas para segmentação, elaboradas no módulo 1, acrescidas com características sobre o nível de profundidade em que se encontram os constituintes na estrutura textual. Em árvores do tipo DTS's, os constituintes identificados são dispostos automaticamente, notadamente, as unidades textuais ocupam lugares pré-determinados em conformidade com a definição da regra de estruturação. Os *segmentos* são identificados como constituintes de 1° nível, ocupando os *nós* principais da árvore. Neste 1° nível podem ser identificados um segmento principal ou vários, de acordo com o tipo de texto.

Na sequência da organização arbórea, os *subsegmentos* são adicionados em outros níveis na árvore em *nós* de 2° e 3° níveis, considerando-se as regras que determinam a posição na DTS. Se houver algum subsegmento em um nível mais profundo na estrutura analisada, por exemplo *nós* de 4° ou 5° níveis, esses constituintes não serão dispostos em *nós* isolados na árvore, permanecendo no interior dos subsegmentos do 3° nível, conforme podemos observar na figura 5.1 na sequência.

No que se refere à avaliação do módulo 2, considerando-se a execução automática das árvores DTS's, optou-se por manter o mesmo padrão utilizado no processo de avaliaçãodo módulo 1, isto é, manual e computacional. A primeira parte da avaliação manual é feita pelo analista nos textos que compõem o corpus *aprendizado*. Nesta etapa, são avaliadas as regras para organização em DTS´s acrescidas com as informações sobre os níveis de profundidade em que se encontram os constituintes, identificados na etapa 1. O analista identifica cada um dos segmentos nos textos e os organiza nas árvores, considerando as regras, conforme apresentamos na tabela 5.7 deste capítulo.

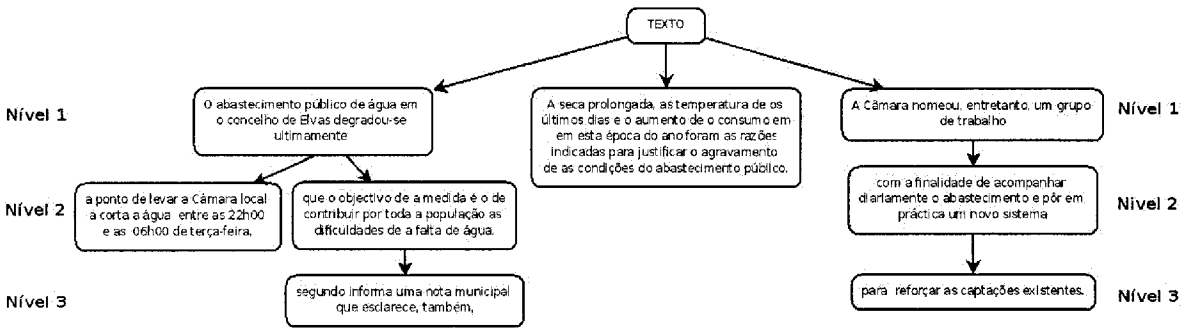

Figura 5.1: A figura apresenta um exemplo de um texto organizado em uma estrutura DTS. O sistema Autema-Dis conserva no 3° nível as estruturas candidatas aos nós de 4° e 5° níveis – publico-19950726-079.

As regras propostas para automatização dos constituintes nas DTS's foram implementadas em prolog e fazem parte das heurísticas destinadas à realização da etapa estrutural. As regras constituidas no sistema AuTema-Dis foram testadas nos textos dos corpora, assim, o conjunto *aprendizado* foi submetido à execução do sistema. Os resultados apresentados pelo sistema foram contrastados com os resultados apresentados na organização das DTS's realizadas manualmente, conforme podemos verificar abaixo na tabela 5.14. As regras identificadas na tabela são utilizadas para organizar unicamente os *subsegmentos* nas DTS's, não sendo estas aplicáveis aos *segmentos*.

Regras e Representação DTS - Avaliação 10 Textos - PE			Correção	
Regras Iniciais	N° de Ocorrências Manual	N° Ocorrências Sistema	n° Profundidade	%
N<:fcl	5	5	5	100%
Advl:pp	42	35	24	69%
Advl:advp	7	7	5	71%
Advl:fcl	1	2	1	50%
Pred:pp	6	3	3	100%
Advl:cu	1	1	1	100%
App:prop	1	1	1	100%
Co:conj-c ('mas')	3	2	2	100%
Pred:np;	2	2	2	100%
App:np	2	2	2	100%
Advl:adv - atemp	6	1	1	100%
Advl:np ou Advl:n	1	2	1	50%
Média	6.4	5.3	4.0	87%

Tabela 5.14: A tabela apresenta os resultados da organização manual e automática dos *subsegmentos* em árvores DTS's, do conjunto *aprendizado*.

Na sequência, o processo de avaliação das regras propostas para organização estrutural em ávores foi realizado no restante dos textos dos corpora, isto é, no conjunto *avaliação/treino*. O procedimento de avaliação é feito manual e automaticamente e em nada difere da avali-

ação realizada nos textos do conjunto *aprendizado*. Os resultados obtidos na apreciação desta etapa são demonstrados na sequência, na tabela 5.15, em que é possível identificar claramente o desempenho do sistema.

Regras e Representação DTS na totalidade dos corpora - 50 textos			Correção	
Regras	Nº de Ocorrências Manual	Nº Ocorrências Sistema	nº Profundidade	%
N<:fcl	38	40	29	73%
Advl:pp	239	205	148	72%
Advl:advp	15	16	15	94%
Advl:fcl	30	34	26	76%
Pred:pp	19	16	13	81%
Advl:cu	7	7	7	100%
App:prop	16	14	12	86%
Advl:acl	8	8	3	38%
Sta:icl	10	13	8	62%
Co:conj-c ('mas')	14	12	11	92%
Pred:np;	13	7	0	0%
App:np	8	10	8	80%
Advl:adv - atemp	52	47	25	53%
Advl:adv - aloc	1	3	0	0%
Advl:np ou Advl:n	6	8	3	38%
Média	31.7	29.3	20.5	70%

Tabela 5.15: A tabela apresenta os resultados da organização manual e automática dos *subsegmentos* em árvores DTS's, do conjunto *avaliação*.

As tabelas 5.14 e 5.15 apresentam equilíbrio se compararmos a organização arbórea manual dos subsegmentos com a organização realizada automaticamente pelo sistema AuTema-Dis. As representações demonstradas em ambas tabelas representam o número de subsegmentos identificados nas árvores *DTS's* a partir da designação de uma regra que identifica o *subsegmento*.

As árvores são geradas a partir da aplicação das regras de segmentação caracterizadas com a informação a respeito dos níveis de profundidade que os subsegmentos podem ocupar na árvore. Desta forma, o sistema identifica inicialmente os subsegmentos que ocupam os *nós* do 2º e 3º níveis, os demais constituintes são reconhecidos pelo sistema como *segmentos* e ocupam os nós de nível 1. Para um segmento ocupar o nível 1 na estrutura DTS's, deverá ser reconhecido pelo sistema, considerando as regras definidas para a identificação dos constituintes do tipo *segmentos*, conforme demonstramos no capítulo 4 na tabela 4.6.

No que se refere à avaliação estatística sistema AuTema-Dis, em executar automaticamente a organização dos constituintes nas DTS's, optou-se pela análise dos seguintes processos:

- a cobertura do sistema em relação à organização dos constituintes dos textos em árvores DTS's;
- a performance do sistema em realizar a organização;
- a correção na distribuição das estruturas nas DTS's.

Neste sentido, as tabelas 5.16 e 5.17 apresentam detalhadamente as medidas estatísticas da execução do sistema em realizar a tarefa de organizar em árvores os segmentos dos textos. A identificação das medidas está pautada nas regras de segmentação e na indicação dos níveis que os contituintes podem ocupar na estrutura arbórea.

Regras de Segmentação - 10 Textos	Precisão	Cobertura	F-Measure
N<:fcl	1.0	1.0	1.0
Advl:pp	0.7	0.6	0.6
Advl:advp	0.7	0.7	0.7
Advl:fcl	0.5	1.0	0.7
Pred:pp	1.0	0.5	0.7
Advl:cu	1.0	1.0	1.0
App:prop	1.0	1.0	1.0
Advl:acl	0.0	0.0	0.0
Sta:icl	0.0	0.0	0.0
Co:conj-c ('mas')	1.0	0.7	0.8
Pred:np;	1.0	1.0	1.0
App:np	1.0	1.0	1.0
Advl:adv - atemp	1.0	0.2	0.3
Advl:adv - aloc	0.0	0.0	0.0
Advl:np ou Advl:n	0.5	1.0	0.7
Macromédia	0.87	0.80	0.70
Micromédia	0.76	0.62	0.69

Tabela 5.16: A tabela apresenta a avaliação estatística com a ocorrência das regras que organizam os constituintes em DTS, realizada no conjunto *aprendizado*.

As tabelas 5.16 e 5.17 apresentam a avaliação do sistema que identifica, classifica e organiza os constituintes em DTS's conforme as regras determinadas. A estatística foi realizada e apresentada separadamente para os dois grupos, *aprendizado* e *avaliação*, para fins de especificações e detalhamento. Na análise das macro e micromédias é possível observar que o sistema apresenta um melhor desempenho para o conjunto *aprendizado/treino* em ambas médias, se compararmos às médias identificadas na totalidade dos corpora, nos conjuntos *avaliação/treino*.

Observou-se um decréscimo na precisão, na cobertura e na F-measure se compararmos as médias obtidas com o resultado das análises realizadas no dez textos do conjunto *apren-*

Regras de Segmentação - 50 Textos	Precisão	Cobertura	F-Measure
N<:fcl	0.73	0.76	0.74
Advl:pp	0.72	0.62	0.67
Advl:advp	0.94	1.00	0.97
Advl:fcl	0.76	0.87	0.81
Pred:pp	0.81	0.68	0.74
Advl:cu	1.00	1.00	1.00
App:prop	0.86	0.75	0.80
Advl:acl	0.38	0.38	0.38
Sta:icl	0.62	0.80	0.70
Co:conj-c ('mas')	0.92	0.79	0.85
Pred:np;	0.00	0.00	0.00
App:np	0.80	1.00	0.89
Advl:adv - atemp	0.53	0.48	0.51
Advl:adv - aloc	0.00	0.00	0.00
Advl:np ou Advl:n	0.38	0.50	0.43
Macromédia	0.73	0.74	0.73
Micromédia	0.70	0.65	0.67

Tabela 5.17: A tabela apresenta a avaliação estatística com a ocorrência das regras que organizam os constituintes em DTS, realizada no conjunto *avaliação*.

dizado/treino com as médias identificadas para a totalidade dos textos no conjunto *avaliação/teste*, as quais encontram-se próximas a 0.03 a 0.05. Nota-se que apenas a macromédia da F-Measure apresentou um aumento de 0.03, não sendo considerado significativo na totalidade da avaliação. Acreditamos que essa baixa nas médias é devida ao aumento no número de ocorrências textuais.

5.1.4 Avaliação do Módulo 3 – Identificação Automática das Relações Retóricas em DTS´s –

A metodologia proposta no capítulo 4 apresenta os módulos que constituem a base do que vem a ser o sistema AuTema-Dis. Na edificação do módulo 3, foi previsto a atribuição automática de algumas relações retóricas entre segmentos e subsegmentos em textos em Língua Portuguesa. A figura 5.2 apresenta um texto organizado em uma estrutura DTS com as relações retóricas devidamente delimitadas entre os constituintes. Devido à questões de ordem semântica, esta etapa foi avaliada de maneira diferenciada em relação as demais etapas implementadas computacionalmente.

No caso do módulo 3, em que avaliamos a atribuição das relações retóricas entre os constituintes dos textos, optou-se por realizar dois tipos de avaliações específicas, são elas:

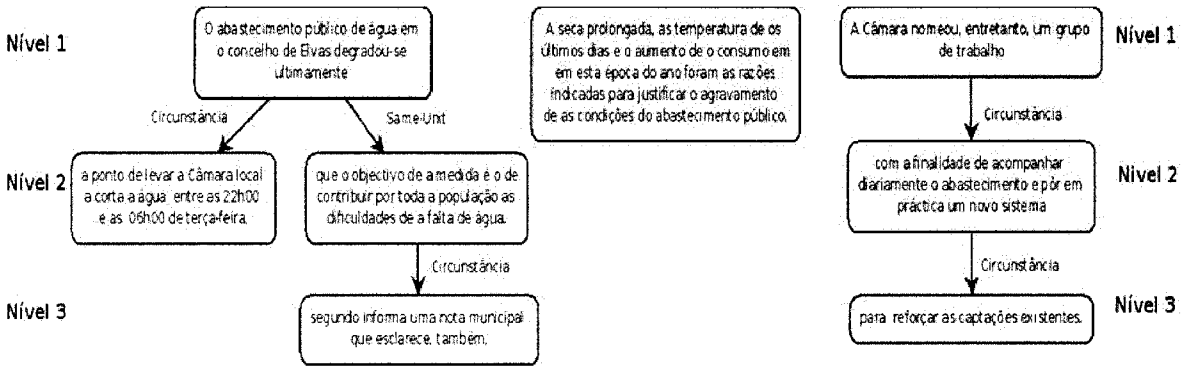

Figura 5.2: A figura apresenta um texto organizado em uma estrutura DTS, as relações retóricas atribuídas entre os constituintes e os níveis que ocupa na estrutura arbórea – publico19950726-079.

- uma avaliação *holística* – em que se verifica qualitativamente a execução do sistema AuTema-Dis, no processo relativo à atribuição das relações retóricas entre os segmentos e subsegmentos;
- uma avaliação *pontual* – em que se avalia quantitativamente a precisão do sistema AuTema-Dis em atribuir corretamente as relações retóricas entre os segmentos e subsegmentos.

No que se refere à avaliação holística, as regras implementadas no sistema AuTema-Dis, que orientam a atribuição automática das relações entre os constituintes dos textos, foram consideradas como informação básica para realizarmos a avaliação. O objetivo em realizar uma análise qualitativa foi verificar a resposta apresentada pelo sistema AuTema-Dis na atribuição automática das relações retóricas entre os segmentos e subsegmentos. A avaliação *holística* foi realizada na totalidade dos corpora comparando-se as análises e automática nos mesmos conjuntos, isto é, *aprendizado/treino* e *avaliação/teste*.

A avaliação foi realizada, inicialmente, respeitando dos resultados apresentados pelo conjunto *aprendizado*, constituído por 10 textos em *PE*. O objetivo foi verificar se o sistema, a partir das regras propostas, seria capaz de executar a tarefa de atribuir automaticamente as relações retóricas entre os segmentos e subsegmentos sem a interferência humana. Ressaltamos que, considerando o tipo de avaliação realizada, estabelecemos prioridades: primeiro foi verificar se o sistema era capaz de fazer a atribuição automática das relações utilizando-se

das regras de segmentação, as quais têm associadas a si algumas das relações retóricas; em um momento posterior do processo de avaliação, o objetivo foi avaliar se o sistema atribuiu corretamente as relações retóricas entre os constituintes.

Considerando a execução automática do sistema AuTema-Dis em atribuir automaticamente algumas das relações retóricas entre os constituintes dos textos, foi possível comparar tais resultados aos resultados da atribuição manual, conforme a tabela 5.18.

Automatização das Relações Retóricas - Corpus Aprendizado		
Textos	Atribuição Manual	Atribuição Automática
Nº 19940101-007	13	11
Nº 19941012-035	6	6
Nº 19941214-076	9	8
Nº 19950519-057	7	7
Nº 19950725-025	6	5
Nº 19950726-079	5	6
Nº 19950916-121	8	6
Nº 19950916-157	10	7
Nº 19950917-041	5	4
Nº 19950814-011	7	1
Total	76	61

Tabela 5.18: A tabela apresenta os resultados da atribuição manual e automática das relações retóricas no conjunto aprendizado.

A partir dos resultados apresentados na tabela 5.18, foi possível evidenciar que o sistema apresentou uma performance adequada, quando contrastada com a análise realizada manualmente. O índice apresentado revela uma pequena diferença entre as duas análises no conjunto *aprendizado/treino*. Todavia essa diferença não constitui uma falta relevante, visto tratar-se de uma avaliação qualitativa – *holística*, em que analisamos se o sistema é capaz de reconhecer e da atribuir relações retóricas entre segmentos e subsegmentos.

No que se refere aos resultados evidenciados no conjunto *avaliação*, observou-se uma pequena diferença entre os resultados do processo manual e o automático, conforme apresentamos nas tabelas 5.19 e 5.20.

Observamos, nas tabelas 5.19 e 5.20, que o sistema AuTema-Dis identifica e atribui um número de relações retóricas muito próximo ao número de relações atribuídas manualmente pelo analista. Os resultados são satisfatórios para a proposta desenvolvida, todavia, fez-se necessário identificar quantas destas relações identificadas e atribuídas automaticamente pelo sistema estão corretas em relação ao que é proposto na análise manual. Os resultados

Automatização das Relações Retóricas - Corpus Avaliação/teste - PE		
Textos	Atribuição Manual	Atribuição Automática
Nº 19940101-007	13	11
Nº 19941012-035	6	6
Nº 19941214-076	9	8
Nº 19950519-057	7	7
Nº 19950725-025	6	5
Nº 19950726-079	5	6
Nº 19950916-121	8	6
Nº 19950916-157	10	7
Nº 19950917-041	5	4
Nº 19950814-011	7	1
Nº 19940504-070	9	6
Nº 19940505-024	5	5
Nº 19940505-071	7	8
Nº 19941911-083	14	14
Nº 19941012-011	12	10
Nº 19941025-045	17	18
Nº 19950416-032	8	6
Nº 19950795-167	13	9
Nº 19950912-022	10	9
Nº 19950924-121	12	13
Nº 19950422-141	9	7
Nº 19950423-011	11	7
Nº 19950629-083	10	10
Nº 19950629-119	12	8
Nº 19951011-139	8	6
Nº 19951011-150	4	4
Nº 19951114-163	7	5
Nº 19951114-169	9	9
Nº 19951220-045	9	7
Nº 19951229-044	11	11

Tabela 5.19: Análise holística da automatização das relações retóricas no conjunto avaliação – Jornal Público.

evidenciados encontram-se dispostos nas tabelas 5.21 a correção do sistema AuTema-Dis na totalidade dos corpora, isto é, considerando os textos do conjunto *aprendizado* e conjunto *avaliação*.

No que concerne à avaliação pontual, realizamos uma apreciação quantitativa dos resultados apresentados pela execução do sistema, conforme apresentamos na tabela 5.21. Verificou-se minuciosamente se as relações retóricas atribuídas pelo sistema AuTema-Dis entre os constituintes estavam corretas em relação à atribuição das relações realizada manualmente. Ressaltamos que, semelhante à avaliação qualitativa, isto é, avaliação holística, a informação básica para este tipo de análise está condicionada ao conjunto de regras, que orientam a

Automatização das Relações Retóricas - Corpus Avaliação/teste - PB		
Textos	Atribuição Manual	Atribuição Automática
Nº FSP950101-011	15	13
Nº FSP950101-032	10	11
Nº FSP950101-054	3	3
Nº FSP950101-084	13	13
Nº FSP950111-014	6	5
Nº FSP950111-026	6	10
Nº FSP950111-034	12	7
Nº FSP950111-036	3	1
Nº FSP950117-048	22	18
Nº FSP950117-074	10	8
Nº FSP940101-132	7	6
Nº FSP940101-124	13	8
Nº FSP940101-107	8	5
Nº FSP940101-102	9	11
Nº FSP940101-095	14	21
Nº FSP940101-092	15	9
Nº FSP940101-085	11	10
Nº FSP940101-079	15	15
Nº FSP940101-074	10	10
Nº FSP940101-066	11	12

Tabela 5.20: Análise holística da automatização das relações retóricas no conjunto avaliação – Jornal Folha de São Paulo.

identificação de uma relação a partir das informações morfo-sintático-semânticas preestabelecidas na própria constituição metodológica das regras.

Com a finalidade de validarmos mais precisamente a execução do sistema AuTema-Dis em atribuir corretamente algumas relações retóricas entre os constituintes, apresentamos uma avaliação estatística do sistema. Na avaliação estatística privilegiou-se a correção das regras atribuídas e não a identificação das ocorrências por texto, para tal, foram consideradas as regras propostas para a realização desta tarefa, conforme podemos observar na tabela 5.22.

Na tabela 5.22 é possível identificar de forma mais pontual os resultados da precisão, cobertura e do desempenho do sistema em atribuir automaticamente as relações retóricas entre os constituintes dos textos analisados. Percebe-se que o AuTema-Dis apresenta satisfatórios resultados no que se refere à precisão e à cobertura na atribuição das relações retóricas entre os constituintes. Na totalidade dos corpora, verificou-se índices semelhantes para as medidas de precisão e cobertura, cerca de 0.60, o que representa um valor positivo, considerando a dificuldade em realizar a tarefa. No que se refere à medida *f-measure*, o sistema apresentou um índice próximo a 0.56, considerado razoável para o tipo de tarefa realizada pelo sistema.

Avaliação da Automatização das Relações Retóricas - 50 textos				
Relações	Regra	N de Rel. Auto.	N de Rel. Corretas	N de Rel. Existentes.
Same-Unit	N<:fcl	40	29	38
Circunstância Genérica	Advl:pp Advl:advp Advl:fcl Pred:pp Advl:cu	200 15 33 16 7	148 15 26 13 7	239 15 30 19 7
Circunstância Apositiva Nome Próprio	App:prop	14	12	16
Avaliação	Advl:acl	8	3	8
Ação	Sta:icl	11	8	10
Oposição Antítese	Co:conj-c ('mas')	12	11	14
Elaboração Circunstância Genérica	Pred:np	6	0	13
Complementação Nominal Apositiva	App:np	8	8	8
Quantificação Temporal	Advl:adv - atemp	47	25	52
Quantificação Locativa	Advl:adv - aloc	3	0	1
Circunstância de Tempo Decorrido	Advl:np ou Advl:n	6	3	6

Tabela 5.21: A tabela apresenta o contraste entre os resultados corretos obtidos pela execução do sistema e os resultados manuais na atribuição das relações retóricas na totalidade dos corpora.

Neste sentido, a partir dos resultado evidenciados através das medidas de avaliação, verificou-se que o AuTema-Dis está capacitado a realizar a tarefa de atribuir relações retóricas entre os segmentos, apesar das suas limitações. Todavia, salientamos, que os resultados obtidos nesta etapa do processamento é satisfatória no âmbito desta proposta.

Avaliação Estatística da Automatização das Relações Retóricas - 50 textos				
Relações	Regra	Precisão	Cobertura	F-Measure
Same-Unit	N<:fcl	0.73	0.76	0.74
Circunstância Genérica	Advl:pp	0.74	0.62	0.67
	Advl:advp	1.00	1.00	1.00
	Advl:fcl	0.79	0.87	0.83
	Pred:pp	0.81	0.68	0.74
	Advl:cu	1.00	1.00	1
Circunstância Apositiva Nome Próprio	App:prop	0.86	0.75	0.8
Avaliação	Advl:acl	0.38	0.38	0.38
Ação	Sta:icl	0.73	0.80	0.76
Oposição/Antítese	Co:conj-c ('mas')	0.92	0.79	0.85
Elaboração - Circunstância Genérica	Pred:np	0	0	0
Complementação Nominal Apositiva	App:np	1.00	1.00	1
Quantificação Temporal	Advl:adv - atemp	0.53	0.48	0.51
Quantificação Locativa	Advl:adv - aloc	0	0	0
Circunstância de Tempo Decorrido	Advl:np ou Advl:n	0.50	0.50	0.50
Média		0.61	0.63	0.56

Tabela 5.22: A tabela apresenta a avaliação estatística do sistema Autema-Dis no processo de atribuição automática das relações retóricas entre os constituintes textuais na totalidade dos corpora.

5.1.5 Avaliação do Módulo 4 – Identificação Automática da Macroestrutura/Macroproposição –

O módulo 4, apresentado e descrito no capítulo anterior, foi elaborado para identificar os segmentos e subsegmentos que contribuem efetivamente na constituição da macroestrutura/macroproposição, bem como, na sua concreta representação. Na essência da sua edificação, a metodologia proposta prevê automatizar a identificação, a seleção e a organização dos constituintes, a fim de gerar uma estrutura representativa do tema do texto sem a interferência humana, conforme pode ser observado nos anexos 4 e 5.

No capítulo anterior, descrevemos detalhadamente a metodologia que é a base o 4° módulo, que constitui uma das etapas do processamento do sistema AuTema-Dis. Trata-se

de um *módulo-resultado*, pois incorpora as informações advindas da realização das tarefas executadas pelo analisador AuTema-dis nas três etapas anteriores. Nesta última etapa do processamento, os resultados são manipulados automaticamente e reorganizados, a fim de que se produza de forma automática uma estrutura representativa do tema do texto analisado, conforme pode ser observado na figura 4.5.

As regras propostas para realizar a identificação, a seleção e a demonstração da estrutura temático-discursiva no módulo 4 fazem parte do conjunto desenvolvido para realizar a 2ª etapa da metodologia, na qual são identificados os segmentos e subsegmentos a comporem as árvores DTS's. Nota-se que no 4º estágio do processamento as regras são acrescidas com informações sobre:

- o nível em que se encontram os constituintes nas estruturas das quais fazem parte.
- o ponto em que uma das macrorregras pode ser aplicada para reorganização/supressão da informação acessória temática da estrutura em questão.

No caso da utilização de macrorregras para auxiliar na elaboração automática da macroestrutura/macroproposição, salientamos que, em nossa abordagem, estamos utilizando apenas as características que envolvem a aplicação da macrorregra *apagamento ou supressão*, sendo esta única das macrorregras selecionada a compor a metodologia, conforme justificamos no capítulo 4.

Para avaliarmos a produção e a apresentação da macroestrutura/macroproposição gerada automaticamente pelo sistema AuTema-Dis, optamos por efetuar dois processos interrelacionados:

1. avaliação das regras para a organização macroestrutural.
2. avaliação do sistema constituído por estas regras.

Inicialmente, verificamos se as regras desenvolvidas para identificação, seleção e organização da macroestrutura/macroproposição estão adequadas a realizarem as tarefas estabelecidas, bem como, se são passíveis de serem implementadas em sistema computacional. Neste sentido, a primeira parte da avaliação é realizada manualmente nos corpora; em um momento posterior, após a validação das regras para execução das tarefas preestabelecidas, procedemos a sua implementação no sistema computacional AuTema-Dis. Assim, realizamos de forma pragmática a avaliação das regras propostas, considerando os resultados da execução do sistema computacional. Dessa forma, realiza-se uma única avaliação, a qual apresenta

dois níveis de resultados:

- avaliação das regras propostas para a identificação, seleção e organização dos constituintes diretamente relacionados ao conteúdo temático;
- avaliação do desempenho do sistema implementado com as regras propostas para produzir automaticamente a macroestrutura/macroproposição do texto analisado.

No tocante ao tipo de avaliação efetuada, as regras foram apreciadas de forma prática a partir da execução do sistema. Assim, verificamos e avaliamos a possibilidade das regras constituirem a base do sistema e; os resultados da sua aplicação no próprio sistema automático desenvolvido exclusivamente para a testagem, isto é, no AuTema-Dis. Em relação à avaliação do resultado da produção da macroestrutura/macroproposição, apresentada ao final da execução do sistema AuTema-Dis, foi realizada uma análise contrastiva entre a identificação da estrutura representativa do tema proposto pelo analista e a identificação apresentada como resultado da execução do sistema AuTem-Dis, a fim de verificar a compatibilidade entre as duas produções.

No caso da *re*-construção das produções das macros foi utilizado o mesmo conjunto de regras tanto para a construção manual, realizada pelo analista, quanto para a construção automática, realizada pelo AuTema-Dis. Em ambas tarefas, empregaram-se as mesmas regras, desenvolvidas para executar essa atividade. Os resultados, os quais transcrevemos na sequência desta seção em tabelas demonstrativas, foram observados utilizando-se como base referencial a proposta apresentada pelo analista em contraste com o resultado gerado sistema.

A realização de todo o processo avaliativo desta 4ª etapa foi feita semelhante à avaliação realizada nas etapas 1, 2 e 3 do processamento. Inicialmente, avaliou-se o conjunto *aprendizado/treino*, constituído por dez textos em Português Europeu *PE*; contabilizados os acertos e identificadas as falhas do sistema, realizou-se a avaliação na totalidade dos corpora, ou seja, os textos representativos do conjunto *avaliação/teste* em *PE* e *PB*, verificando-se os resultados apresentados pela execução do sistema nos 50 exemplares.

Conforme apresentamos, os textos produzidos pelo sistema, representativos do conteúdo temático, foram avaliados em termos de correção, a partir da proposta apresentada pelo analista. As incorreções observadas nos resultados produzidos automaticamente pelo AuTema-Dis foram minuciosamente avaliadas no sentido de determinarmos qual a causa ou elemento responsável pelo erro. Na busca pelo reconhecimento do problema que compromete

o adequado desempenho do AuTema-Dis, identificamos:

- se a falha teve origem em uma regra incorretamente elaborada ou mal empregada/selecionada pelo sistema na sua execução;
- se o erro fora produzido da primeira etapa do processamento, a qual classifica os segmentos e subsegmentos a partir da análise automática do *Palavras*.

A necessidade em identificarmos a origem do *erro* na produção automática da macroestrutura/macroproposição justifica-se no sentido de que buscamos uma execução do sistema AuTema-Dis próxima ao que possa ser considerado aceitável como uma produção coesa e coerente. Na sequência desta seção, apresentamos as tabelas 5.23, 5.24 e 5.25 com as avaliações estatísticas da execução do sistema AuTema-Dis na produção automática da macroestrutura/macroproposição nos textos dos corpora.

Em termos concretos, a avaliação do módulo 4 foi realizada de forma contrastiva, tomando-se como referência as macroestruturas/macroproposições apresentadas pelo analista para os textos dos corpora. Como se previu, não existe uma identificação totalmente perfeita entre ambas ocorrências, visto tratar-se de uma produção humana, que lida com conhecimentos extra-linguísticos, que envolvem conhecimento de mundo nos níveis conceituais e pragmáticos; e uma produção automática, que lida apenas com informações de ordem linguística, exclusivamente do meio textual.

Para tornarmos o processo de avaliação da produção automática da macroestrutura/macroproposição mais objetivo e pontual, elaboramos uma estratégia para nos afastarmos as questões subjetivas da comparação entre a a análise da produção manual da macroestrutura e à produção automática, visto que não existe paridade entre as duas produções. Desta forma, atribuídos um padrão em termos percentuais para cada *erro* encontrado na produção automática; ou seja, um erro equivale a noventa porcento (90%) de acerto na produção e assim sucessivamente até a totalidade de dez (10) erros, o que invalida a macroestrutura/macroproposição produzida pelo sistema.

Conforme mencionamos, os *erros* podem ser da ordem do analisador sintático *Palavras*, que envia a análise com problemas para o módulo 2 ou; da ordem das regras de segmentação, que apresentam um grau de generalização muito profunda na seleção dos constituintes. Outro ponto a ser salientado é que podem ocorrer situações em que existam erros ou incoerências entre os resultados da identificação manual e os resultados da identificação automática que não podem ser validados em termos percentuais pois, apesar de serem observados no âmbito da estrutura produzida pelo sistema, não comprometem do reconhecimento do tema.

Todavia, procuramos no processo avaliativo, abster-nos da subjetividade inerente à análise realizada pelo analista.

Desta forma, a análise estatística foi realizada em termos de correção, ou seja, certo e errado, não sendo possível avaliar a precisão, a cobertura e *f-measure* no que diz respeito à execução da 4ª etapa do sistema AuTema-Dis. Como se trata de uma avaliação com características próximas à subjetividade, só foi possível identificar os erros e o percentual da correção na estrutura produzida pelo sistema, considerando as estruturas propostas analista como sugestão para uma macroestrutura/macroproposição correta.

Conforme mencionamos, se o erro apresentado no interior da estrutura não compromete o seu conteúdo, a sua informação e, se a estrutura gerada reproduz de forma satisfatória o que está disposto na superfície do texto analisado, a macroestrutura/macroproposição apresentada automaticamente será considerada adequada na sua totalidade, isto é, (100%) correta. Ressaltamos esta possibilidade, pois identificamos este tipo de situação em algumas das ocorrências estruturais geradas pelo sistema. Observou-se, por exemplo, que o AuTema-Dis apresenta dificuldades ou, em algumas situações, não consegue classificar um constituinte do tipo *sigla*, bem como, determinadas construções/informações que se encontram nos parênteses, colchetes e chaves. Este fato ocorre devido a duas situações específicas: por estar faltando uma regra adequada para contemplar estes casos; ou porque o analisador *palavras*, quando realizou a análise no âmbito da primeira etapa do processamento do AuTema, atribuiu uma simbologia desconhecida ou diferenciada, a qual não está determinada no conjunto das regras preestabelecidas.

Todavia, ao analisarmos os dados apresentados, verificamos que os resultados da avaliação da produção automática da macroestrutura/macroproposição apresentou índices satisfatórios na correta produção das estruturas na totalidade dos corpora. Os dados podem ser evidenciados nas tabelas 5.23, 5.24 e 5.25.

Verificou-se, com base nos resultados demonstrados nas tabelas, que o sistema AuTema-Dis apresentou uma performance muito satisfatória no que se refere ao percentual de correção no processo de geração automática da macroestrutura/macroproposição. A avaliação foi realizada nos mesmos corpora que foram utilizados nas avaliações dos outros módulos. Observou-se que as médias de correção na produção das macros é superior a 70%, índice considerado positivo para uma primeira versão do sistema.

Percebe-se, ao avaliar atentamente os resultados, que o grupo constituído pelos 10 textos, os quais constituem o conjunto *aprendizado/treino*, apresentou uma correção (100%) em

Avaliação da Macroestrutura/ Macroproposição Automática Conjunto Aprendizado - Corpus Jornal Público - 1994/1995				
Textos	Correção	Nº de Erros	Erro Palavras	Erro Regras
Nº 19940101-007	100%	0	0	0
Nº 19941012-035	80%	2	1	1
Nº 19941214-076	100%	0	0	0
Nº 19950519-057	80%	2	1	1
Nº 19950725-025	60%	1	1	0
Nº 19950726-079	100%	0	0	0
Nº 19950916-121	90%	1	0	1
Nº 19950916-157	100%	0	0	0
Nº 19950917-041	100%	0	1	0
Nº 19950814-011	80%	1	1	0
Média	89	0.7	0.5	0.3

Tabela 5.23: A tabela apresenta a estatística relativa ao resultado da geração da macroestrutura/macroproposição no conjunto aprendizado.

Avaliação da Macroestrutura/Macroproposição Automática Conjunto Avaliação - Corpus Jornal Público - 1994/1995				
Textos	Correção	Nº de Erros	Erro Palavras	Erro Regras
Nº 19940504-070	90%	1	0	1
Nº 19940505-024	80%	1	1	1
Nº 19940505-071	80%	2	1	1
Nº 19941911-083	70%	3	2	1
Nº 19941012-011	100%	0	0	0
Nº 19941025-045	70%	2	1	1
Nº 19950416-032	60%	2	2	0
Nº 19950795-167	70%	3	3	0
Nº 19950912-022	60%	2	2	0
Nº 19950924-121	50%	3	2	1
Nº 19950422-141	60%	1	0	1
Nº 19950423-011	50%	4	3	1
Nº 19950629-083	100%	0	0	0
Nº 19950629-119	70%	3	2	1
Nº 19951011-139	90%	1	0	1
Nº 19951011-150	100%	0	0	0
Nº 19951114-163	90%	1	0	1
Nº 19951114-169	80%	2	0	2
Nº 19951220-045	100%	1	1	0
Nº 19951229-044	100%	0	0	0
Média	78.5%	1.6	1	0.65

Tabela 5.24: A tabela apresenta a estatística relativa ao resultado da geração da macroestrutura/macroproposição no conjunto avaliação referente ao Jornal Público.

Avaliação da Macroestrutura/Macroproposição Automática Conjunto Avaliação - Corpus Folha de São Paulo - 1994/1995				
Textos	Correção	N° de Erros	Erro Palavras	Erro Regras
N° FSP950101-011	90%	1	0	1
N° FSP950101-032	80%	1	1	0
N° FSP950101-054	100%	0	0	0
N° FSP950101-084	50%	5	5	0
N° FSP950111-014	100%	0	0	0
N° FSP950111-026	100%	2	2	0
N° FSP950111-034	100%	0	0	0
N° FSP950111-036	100%	0	0	0
N° FSP950117-048	90%	2	1	1
N° FSP950117-074	100%	0	0	0
N° FSP940101-132	80%	2	2	0
N° FSP940101-124	90%	2	2	0
N° FSP940101-107	100%	0	0	0
N° FSP940101-102	100%	0	0	0
N° FSP940101-095	100%	0	0	0
N° FSP940101-092	80%	3	1	2
N° FSP940101-085	100%	2	2	0
N° FSP940101-079	90%	1	1	0
N° FSP940101-074	100%	0	0	0
N° FSP940101-066	100%	1	1	0
Média	92.5%	1.9	0.9	0.2

Tabela 5.25: A tabela apresenta a estatística relativa ao resultado da geração da macroestrutura/macroproposição no conjunto avaliação/teste referente ao Jornal Folha de São Paulo.

metade dos textos do conjunto, o que confere uma média global de (78%), muito acima do esperado. Nos dois grupos que compreendem o conjunto *avaliação/teste*, os 20 textos em Português Brasileiro apresentam uma correção de (100%), mais da metade dos textos, o que representa (92%) de correção neste grupo.

O grupo constituído pelos 20 textos em Português Europeu apresenta um resultado inferior ao que foi apresentado no conjunto do *PB*; no conjunto em *PE* obteve-se o índice de correção de (100%) em apenas 3 textos do total de 20; com este índice, a média de correção das macros, para este conjunto, ficou próximo a (78,5%). Acreditamos que a justificativa para este índice de correção inferior aos demais está relacionada à construção das estruturas frasais que compõem os *PE*. No entanto, não podemos validar a hipótese apresentada, pois esta possibilidade não foi investigada no âmbito desta tese.

Para finalizar a descrição da avaliação do módulo 4, salientamos que não existe uma rees-

critura na apresentação automática do tema do texto, mas sim, existe uma seleção e *(re)*organização dos segmentos e subsegmentos que detém a informação relacionada à temática discursiva. É consensual que não existe, até o momento, a possibilidade de que um sistema automático se aproprie dos constituintes textuais, absorvendo o seu conteúdo semântico, para apresentar uma estrutura totalmente *remodelada*. Essa possibilidade é descartada em nosso sistema por motivos elementares. Todavia, apesar do sistema não gerar uma estrutura *remodelada* em termos de escrita, o que é apresentado pelo sistema AuTema-Dis apresenta condições mínimas para a compreensão do que é tratado no âmbito do texto analisado.

5.2 Avaliação Geral: metodologia – implementação

No tocante à avaliação da metodologia proposta para a análise textual e identificação da temática discursiva, foi possível evidenciar que o AuTema-Dis apresenta resultados satisfatórios e equilibrados na realização de todas as etapas constituintes da análise textual, a partir do processo de sistematização dos módulos de análise.

Os resultados evidenciados a partir das etapas realizadas pelo sistema AuTema-Dis demonstram que a elaboração e a realização da metodologia, que constitui a base do sistema computacional, está constituída adequadamente, conforme podemos observar no anexo 5. Os dados estatísticos, evidenciados na avaliação do desempenho do sistema em realizar cada uma das etapas propostas, possibilitou-nos concluir que existe equilíbrio quanto aos resultados, visto que, as etapas utilizam reciprocamente os resultados umas das outras, para que o processamento ocorra na íntegra e de maneira satisfatória, o que foi evidenciado nos resultados das análises.

Justificamos a avaliação como positiva, a partir dos números evidenciados nas estatísticas realizadas para avaliar a proposta metodológica, constituída pelas regras, bem como, pelo desempenho do sistema em executá-las sistematicamente. Um ponto a ressaltar é que caso uma das etapas de análise apresente um resultado pouco satisfatório na execução de uma tarefa, esse resultado pode comprometer, na sequência do processamento, o resultado da etapa seguinte. Um exemplo que sustenta o parecer apresentado é o que ocorre a partir de uma má identificação e classificação dos constituintes textuais na 1ª etapa, a qual poderá comprometer a organização nas árvores, a 2ª etapa do processamento, e assim, sucessivamente até o término da análise. Outrossim, vale ressaltar que, por se tratar de uma metodologia sequencial em que, por exemplo, a realização da 3ª etapa depende dos resultados do processamento das duas etapas anteriores, mesmo se o utilizador do sistema não estiver interessado nos resultados das outras etapas, tais resultados estão disponíveis, pois fazem

parte do encadeamento do sistema.

O desempenho do sistema AuTema-Dis em realizar a análise textual e identificação da temática discursiva automaticamente, sem a interferência humana, pode ser relacionado e comparado ao desempenho apresentado pelos tradutores automáticos e tradutores online. Algumas vezes, as estruturas não recebem o tratamento adequado na apresentação do conteúdo, o que não compromete necessariamente o seu entendimento por parte do seu utilizador/leitor.

5.3 Resumo do Capítulo

Neste capítulo, apresentamos a avaliação da metodologia proposta para análise automática discursiva, devidamente implementada em sistema computacional, denominado AuTema-Dis. A avaliação apresentada foi realizada sequencialmente em conformidade com a execução das etapas do processamento do sistema. Trata-se, portanto, de uma avaliação modular. Assim, descrevemos a análise realizada, bem como, os resultados obtidos com a conclusão do processamento e a avaliação estatística, a qual representa a precisão, a cobertura e a performance do sistema.

Na sequência desta tese, apresentamos o capítulo com as conclusões, os trabalhos complementares, os avanços e as melhorias que podem ser feitas no âmbito da metodologia e da própria execução do sistema do sistema AuTema-Dis.

Capítulo 6

Conclusões

O trabalho descrito faz parte da investigação realizada no âmbito do Doutoramento em Informática, na área de Linguísitca Computacional. Trata-se, como apresentamos inicialmente de um estudo interdisciplinar que coaduna duas áreas distintas, a Informática e a Linguística. A investigação realizada teve como objetivo apresentar uma nova proposta metodológica para análise discursiva, cujo processo pudesse servir de base a uma arquitetura computacional para realizar uma análise automática completa texto/discurso, sem a intervenção humana. Assim, a proposta foi organizada em duas partes distintas:

1. Metodologia – elaborou-se um constructo teórico-metodológico para análise textual e identificação automática da temática discursiva em textos em Português;

2. Arquitetura – construiu-se uma estrutura modular destinada à implementação da metodologia em sistema computacional, identificado, nesta investigação, acrônimo AuTema-Dis – automatização da temática discursiva.

6.1 AuTema-Dis - uma metodologia em sistema

A metodologia proposta está relacionada à representação da temática discursiva e foi elaborada em módulos autônomos para a realização das análises nos textos e apresentação dos resultados na sua totalidade. Na elaboração da metodologia optou-se pela objetividade e simplicidade na execução das etapas previstas, as quais seguem uma organização e ordenação pré-determinada no processo de realização das tarefas, ou seja, na análise modular dos textos.

Conforme apresentamos nos capítulos 4 e 5, cada uma das etapas da metodologia foi desenvolvida tendo em vista uma contrapartida computacional, o que equivale, em termos

representacionais, que cada um dos módulos da metodologia compreenda uma etapa do processamento no sistema automático. Neste sentido, conforme demonstramos, a metodologia foi estruturada em quatro módulos, em que cada um destes módulos apresenta autonomia na sua realização e apresentação de resultados.

Todavia, salientamos que os resultados obtidos a cada etapa executada pelo sistema servem de base, ou seja, detém a informação necessária à execução da etapa seguinte de análise. Assim, a metodologia deve ser compreendida como um proposta modular autônoma e inter-relacionar, constituída por 4 módulos, os quais edificam a arquitetura computacional AuTema-Dis, para análise da temática discursiva.

6.1.1 Ferramenta Modular

A metodologia é identificada e definida como uma *ferramenta modular*, a qual se destina a realizar diferentes tipos de análises textuais, em conformidade com o módulo/etapa requisitada pelo usuário. Na metodologia foram edificados 4 módulos básicos para realizarem análises textuais específicas e autônomas e inter-relacionadas, são elas:

- análise automática de textos em *PB* e *PE*, reconhecimento dos constituintes mínimos significativos (segmentos e subsegmentos) nas estruturas que compõem a superfície textual;
- análise dos constituintes textuais, classificação quanto ao papel desempenhado na estrutura textual e organização arbórea - DTS's;
- análise dos constituintes textuais dispostos nas DTS's e atribuição automática das relações retóricas entre eles;
- análise dos resultados dos três módulos anteriores, tratamento dos resultados e produção automática da macroestrutura/macroproposição textual.

Cada módulo executa uma tarefa única, a qual tem como objetivo apresentar um resultado exclusivo, a ser considerado ou não para a realização do(s) módulo(s) seguintes. A forma como propusemos a sistemática de análise, os resultados produzidos podem ser analisados isoladamente ou na totalidade da execução das etapas, conforme a determinação e necessidade do usuário. Em casos específicos, o usuário poderá determinar os resultados que deseja, optando pela etapa que desejar, todavia, o sistema executará automaticamente as etapas necessárias para cumprir a solicitação, isto é, dependendo da necessidade do utilizador, o

sistema terá que executar algumas etapas precedentes à etapa que deseja como resultado.

Neste sentido, a metodologia é caracterizada e definida como uma *ferramenta modular* autônoma, justificando-se, desta forma, a completa realização de cada uma das etapas na constituição do resultado final, a representação temática do texto analisado. Além disso, a *independência* na execução de cada módulo e a inter-relação com os demais possibilita que novas tarefas possam ser agregadas aos módulos já existentes, para complementar e ampliar os resultados apresentados. A arquitetura AuTema-Dis prevê na sua origem que possam ser agregados novos módulos de processamento, conforme a evolução das investigações na análise automática de discurso.

6.1.2 Metodologia AuTema-Dis: contribuições *Lato Sensu*

A metodologia desenvolvida para análise discursiva foi idealizada com a finalidade de que cada um dos módulos, implementado em sistema computacional, pudesse ser facilmente ampliado e reformulado, conforme os resultados de novos estudos nesta área de análise discursiva de cada um dos módulos de execução das tarefas, ou incluindo-se nas possibilidades de ampliação da metodologia/sistema a criação/junção de outros módulos aos já existentes. Assim, consideramos a organização metodológica com uma das importantes colaborações desta investigação.

No início desta tese apresentamos algumas motivações que nos conduziram à realização deste trabalho, as quais direcionaram a organização da metodologia, notadamente, vislumbramos as contribuições que a realização da pesquisa pudesse promover não somente em termos acadêmicos, mas também, contribuições externas, as quais pudessem contribuir fora do escopo científico. Neste sentido, a proposta AuTema-Dis pode contribuir em *Lato Sensu*:

- sistemas de busca automática de resumos de textos na web ou em banco de dados sem limitações aos usuários;
- sistemas de recuperação de informação e sistemas de pergunta/resposta, por exemplo na utilização de descrições em que o usuário busca uma resposta direta e simples;
- sistemas de produção de resumo automático em textos, dinamizando a busca pelo tópico pesquisado;
- sistemas de busca em extratos textuais, resgate de segmentos específicos de textos;

6.1.3 Metodologia AuTema-Dis: contribuições *Stricto Sensu*

As principais contribuições desta investigação estão relacionadas a própria constituição da metodologia do tipo modular, desenvolvida para análise da temática discursiva e sua implementação em sistema automático, que executa todo processamento textual sem a interferência humana. Desta forma, a metodologia e implementação pode contribuir, em *Stricto Sensu*, nos seguintes processos:

- identificação automática dos constituintes textuais e delimitação de seus limites e fronteiras;
- identificação e classificação dos constituintes textuais em relação ao papel temático desempenhado no discurso;
- padronização no processo automático de segmentação textual, eliminando parcialmente problema da granularidade na identificação das estruturas, bem como, sua extensão;
- inclusão dos níveis de profundidade em que se encontram os constituintes textuais na estrutura e a sua caracterização em relação ao compromisso que desempenham em relação à temática discursiva;
- automatização na identificação e na atribuição de algumas relações retóricas entre os constituintes nos textos, independente de frases-pistas, de marcadores discursivos, de critérios de pontuação e elementos exclusivamente sintáticos;
- manipulação dos recursos micro e macrotemáticos como instrumentos na reestruturação da coerência macropropositiva/macroestrutual;
- apresentação de novas relações retóricas, condicionadas pelos critérios propostos na base da *RST*;
- elaboração automática da macroestrutura/macroproposição textual sem a interferência humana a não ser no momento de selecionar o texto a ser analisado;
- automatização da temática discursiva em produções escritas em *PE* e *PB*.

Assim, elaboramos a proposta metodológica para automatização da temática discursiva que apresentamos nesta tese e a sua implementação em sistema automático - AuTema-Dis.

6.1.4 Metodologia AuTema-Dis: limitações

A metodologia AuTema- Dis apresenta algumas limitações e pontos a serem reestruturados, como é previsível quando se constrói uma nova proposta teórica. No que diz respeito à metodologia e sua execução no sistema automático apresentados nesta tese, evidenciamos as seguintes limitações:

- tratamento adequado aos títulos, quando não se encontram pontuados, como é o caso dos textos jornalísticos.

- tratamento adequado ao elemento verbal, considerando-se a construção de uma ontologia robusta para os verbos, a fim de determinarmos mais formalmente os complementos e delimitarmos de forma mais precisa se podem ou não serem retirados ou omitidos no momento da constituição da macroestrutura/macroproposição.

- identificação de um conjunto muito restrito de Relações Retóricas e Relações Estruturais, ponderando-se a possibilidade de desenvolver regras mais específicas para uma identificação mais conceitual.

- produção pouco refinada para a macroproposição/macroestrutura; buscar equilíbrio e harmonia conceitual, a fim de que o sistema seja capaz de produzir de forma mais elaborada tais estruturas.

- resultado da análise automática do *PALAVRAS* – o analisador em algumas ocasiões apresenta falhas na análise e geração de resultados, fato que e compromete os resultados das etapas realizadas pelo AuTema-Dis.

- tratamento automático do AuTema-Dis encontra-se, no momento, limitado ao texto jornalístico do tipo *notícia*, não há testes com outras tipologias textuais.

- reconhecimento dos pronomes anafóricos e indexação nominal. A metodologia AuTema-Dis não prevê a identificação dos anafóricos e o seu correspondente nominal.

- tratamento dos marcadores discursivos a partir das relações retóricas instituídas por esses elementos linguísticos.

Apresentamos as limitações da metodologia e do sistema AuTem-Dis, acreditamos na possibilidade de ajustarmos os itens apresentados na busca por um melhor desempenho do sistema.

6.1.5 Dificuldades Linguístico-Computacionais Equacionadas pelo AuTema-Dis

Na análise da literatura na área da análise automática de discurso, verificamos algumas dificuldades da ordem linguístico-computacional em a adequar conceitos e metodologia à implementação em sistema computacional. Muitas vezes a relação entre metodologia e terminologia não ocorre de forma direta ou linear. Neste sentido, evidenciamos que a metodologia implementada foi capaz de equacionar algumas dessas dificuldades apontadas na literatura, conforme podemos evidenciar:

- Identificação e divisão automática das unidades de análise textuais.
- Granularidade na segmentação dos constituintes textuais.
- A identificação automática das unidades mínimas de segmentação - Edu's (segmentos principais e subsegmentos) e categorização destes elementos quanto ao papel que representam na estrutura.
- Identificação e Atribuição automática de algumas das relações retóricas entre os constituintes textuais.
- Apresentação automática de uma estrutura representativa do tema do texto com base em macroproposições localizadas e não em índices lexicais.

6.2 Avaliação dos Resultados

A metodologia para análise discursiva automática – AuTema-Dis foi desenvolvida e avaliada de forma modular, conforme o processo de análise executado pelo sistema. A proposta de análise elaborada e descrita nesta tese é inédita, o que dificultou o processo de avaliação, pela dificuldade em estabelecermos contraste com outros sistemas que também realizam, de certa forma, uma análise discursiva automática.

No caso específico do analisador AuTema-Dis, este percorre toda a estrutura discursiva, analisando os constituintes textuais e as relações que se estabelecem entre eles na constituição da temática discursiva. Todo o processo é realizado sem a interferência humana na apresentação dos resultados em cada uma das etapas.

Neste sentido, não foi possível comparar os resultados obtidos em cada uma das etapas processadas e avaliadas com os resultados de outros sistemas. Todavia, realizamos avaliação

da precisão, cobertura e performance do sistema considerando como referencial os resultados apresentados por um analista na execução da mesma atividade.

De forma geral o sistema respondeu de forma satisfatória à execução das tarefas propostas em cada um dos módulos, conforme apresentamos no capítulo 5 desta tese.

6.3 Trabalhos Futuros

No que se refere à perspectiva de trabalhos futuros a partir da metodologia proposta, acreditamos na possibilidade de ampliar e aprofundar o sistema AuTema-Dis, no que se refere ao processamento automático da temática discursiva. Inicialmente, poder-se-ia trabalhar com o refino do processamento de cada um dos módulos do sistema, buscando aprimorar os resultados existentes já validados estatisticamente.

Como mencionamos, a Metodologia AuTema-Dis prevê na sua origem a possibilidade de expansão e complementação dos módulos já existentes, bem com, a agregação de novos módulos no processo de análise. A possibilidade de expansão dos módulos, que são constitutivos da metodologia, está relacionada à busca mais precisa e pontual para os resultados finais da execução de cada um dos módulos/etapas da metodologia/sistema. Neste sentido, vale ressaltar que uma expansão do sistema poderia estar relacionada:

- à categorizaçao mais específica e limitativa para os constituintes verbais, a fim de delimitarmos mais pontualmente os complementos, bem como, nível que podem ocupar na estrutura arbórea, relacionando essa categorização à relevância da identificação temática;
- ao desenvolvimento uma etapa para identificação dos constituintes nominais indexados a uma expressão anafórica ou catafórica. Esta especificação seria importante para classificar o nível em que se encontra um determinado constituinte e a sua relevância na macroprosição local e global;
- à elaboração de uma metodologia que indexasse os marcadores discursivos às relações retóricas, o que possibilitaria determinar a relação semântica existente entre os constituintes.

Além das possibilidades específicas à constituição dos módulos, seria pertinente testarmos o sistema AuTema-Dis com outras tipologias textuais e com outros idiomas de origem românica. Neste sentido, realizamos testes com a tradução de textos escritos em Língua Espanhola e

Língua Galega, os quais fazem parte do banco de dados da Universidade de Compostela – ES. No entanto, não fora realizado um levantamento suficientemente exaustivo que pudesse constituir o processo de avaliação, bem como, validar o desempenho do sistema AuTema-Dis em uma língua diferente do Português. Todavia, de uma forma preliminar, os resultados evidenciados pela execução do sistema nos textos em Espanhol e Galego revelaram-se adequados.

6.4 Considerações Finais

A proposta metodológica para análise automática discursiva e a sua implementação em sistema computacional apresentada no âmbito desta tese, representou um desafio, devido à interdisciplinariedade envolvida no processo. Verificou-se a necessidade e a dificuldade em compor a metodologia que realizasse exaustivamente a análise textual completa, sem a interferência humana e que, além disso, esta metodologia fosse passível à implementação em sistema computacional.

Neste sentido, o resultado do desafio inicial foi realizado com sucesso. A metodologia proposta foi implementada em um sistema computacional, AuTema-Dis, e encontra-se disponível à comunidade, como uma ferramenta para análise discursiva no seguinte endereço: http://abc.di.uevora.pt/ lr/interfaceweb/index.php.

Apêndices

Apêndice A

Os Corpora

A.1 Os Corpora: Aprendizado e Avaliação

A.1.1 Corpus Jornal Publico 1994/1995 – Conjunto Aprendizado/Treino –

1. Público - 19940101–007
   Morreu "Sonny" Constanzo. Dominic "Sonny" Constanzo, que acompanhou, com o seu trombone, cantores como Ella Fitzgerald e Tony Bennett, morreu quinta-feira, em New Haven, no estado americano do Connecticut, aos 61 anos, depois de um transplante cardíaco. Ao longo da sua carreira tocou com o clarinetista Woody Herman, o trompetista Thad Jones, o baterista Mel Lewis e o cantor Clark Terry. Durante muito tempo acompanhou a vocalista Rosemary Clooney, à frente da sua grande orquestra. Mas apenas em 1992 conseguiu fazer a primeira gravação para uma grande etiqueta, no caso a Stash.

2. Público - 19941012–035.txt
   Investidores institucionais. Os operadores detectaram, durante a sessão de ontem da bolsa de Londres, que investidores institucionais britânicos e norteamericanos estiveram particularmente activos, o que se reflectiu no fecho em alta desta praça. Com efeito, o FTSE, o índice da Bolsa de Londres, recuperou 40,7 pontos, face a segunda-feira, para fechar nos 3073 pontos.

3. Público - 19941214–076 DIR

   Manifestação estudantil fracassa em Almada. Escassas dezenas de estudantes participaram ontem numa manifestação estudantil em Almada, destinada a protestar contra a detenção de dois alunos há cinco dias. Um dos jovens, aluno do 11° ano, havia

sido detido pela PSP de Almada durante uma outra manifestação. Um colega que se dirigiu à esquadra, para saber do seu paradeiro, acabou por ser obrigado a lá ficar também toda a noite. Alegada injúria e desobediência aos agentes estiveram na base das detenções. O protesto de ontem, que foi organizado pelos dirigentes estudantis da escola secundária do Monte da Caparica, acabou por não passar de um reduzido grupo de jovens reunidos na Praça São João Baptista – apesar dos anunciados convites aos líderes estudantis das outras escolas da Grande Lisboa e do resto do país. «Não houve 'aderência'», lamentava um aluno depois de finda a manifestação.

4. Público - 19950519–057
   Detectar a doença de Alzheimer. Um estudo finlandês realizado pela Universidade de Kuopio conclui que é possível detectar a doença de Alzheimer cinco a dez anos mais cedo do que se pensava até aqui. Kalevi Pyorala, líder do estudo, afirma que os médicos em geral diagnosticam a doença por volta dos 70 anos, mas que é possível em 35 a 37 por cento dos casos distinguir as perturbações de memória típicas da doença das provocadas por outras causas. O estudo de Pyorala, que envolveu 1100 doentes dos 65 aos 78 anos, foi publicado na revista "Neurology".

5. Público - 19950725–025
   Ligeira queda. As acções cotadas na praça australiana terminaram em baixa, após a tomada de mais valias em virtude das subidas verificadas nas sessões anteriores, disseram os operadores. O índice da Bolsa de Sidney, o All Ordinaries, perdeu 2,8 pontos durante a sessão e fechou nos 2108,7 pontos, menos 0,13 por cento face ao valor de fecho de sexta-feira.

6. Público - 19950726–079
   Elvas, Água só de dia. O abastecimento público de água no concelho de Elvas degradou-se ultimamente a ponto de levar a Câmara local a cortar a água entre as 22h30 e as 06h00 de terça-feira, segundo informa uma nota municipal que esclarece, também, que o objectivo da medida é o de "distribuir por toda a população as dificuldades da falta de água". A seca prolongada, as temperaturas dos últimos dias e o aumento do consumo nesta época do ano foram as razões indicadas para justificar o agravamento das condições do abastecimento público. A câmara nomeou, entretanto, um grupo de trabalho com a finalidade de acompanhar diariamente o abastecimento e pôr em prática um novo sistema para reforçar as captações existentes.

7. Público - 19950916–121
   Recorde adiado. Ainda não foi nesta semana que caiu o recorde absoluto da Bolsa de Londres. A maior subida registou-se na sessão de quarta-feira, mas essa tendência não teve sustentação nas duas sessões subsequentes. As expectativas de uma descida nas taxas de juro britânicas e norteamericanas, juntamente com os bons resultados de Wall Street sustentaram esta subida. Agora os "dealers" esperam que o mercado consolide, antes de tentar um novo assalto ao recorde.

8. Público 19950916–157
   Botins e saltos altos. Em matéria de calçado o castanho e preto serão as cores mais usadas neste Outono/Inverno. Botas e botins práticos, mocassins e sapatos baixos numa linha bastante simples são as tendências para o dia-adia. Os sapatos clássicos têm formas arredondadas ou bicudas. Para a noite, muitas tiras que apertam o tornozelo e outras que cruzam no peito do pé. Os saltos são altos tornando-se o complemento de toda uma elegância. Quanto aos materiais, dominam as vitelas italianas, "nobucks" e crutes acamurçados. Estas são as propostas da Hera para esta estação, que pode encontrar em muitas das sapatarias do país ou nas lojas Hera, em Lisboa (Centro Comercial Amoreiras), Estoril (CascaisShopping), Coimbra (Coimbra Shopping) e Faro (Trav. Rebelo da Silva).

9. Público - 19950917–041
   Cabo Verde. Autarcas portugueses no combate à cólera. Na sequência de um pedido da Câmara Municipal da Cidade da Praia, a Associação Nacional de Municípios está a lançar uma campanha de solidariedade do poder local português com os autarcas cabo-verdianos. O objectivo é ajudar a combater a epidemia de cólera que há vários meses assola Cabo Verde. De acordo com um comunicado da Associação Nacional de Municípios, vai ser enviada de imediato para este país uma primeira remessa de medicamentos e materiais de desinfecção e de prevenção. "A iniciativa tem vindo a encontrar a melhor receptividade nos municípios portugueses", refere a Associação Nacional de Municípios.

10. Público - 19950814–011
    Marcel Moussy. O cineasta francês Marcel Moussy, argumentista dos filmes "Os 400 Golpes", de François Truffaut e "Paris está a arder?", de René Clément, morreu em Caen, França, com 71 anos de idade, de doença prolongada, divulgou ontem a sua família. Marcel Moussy realizou várias obras de ficção para o cinema e para a televisão. É também autor de várias peças de teatro e romances. A sua última obra de ficção, «Un Parfum d'Absinthe», publicada em 1990, foi galardoada com o Prémio Albert Camus.

A.1.2 Corpus Jornal Publico 1994/1995 – Conjunto Avaliação/Teste –

1. Público - 19940504–070
   Montijo adere ao PER A CÂMARA MUNICIPAL DO Montijo assina hoje com o Instituto Nacional de Habitação e o Instituto de Gestão e Alienação do Património Habitacional do Estado (INH e IGAPHE), o acordo de adesão ao Programa Especial de Realojamento (PER), um "pacote" governamental que visa a erradicação das barracas até ao ano 2000. A cerimónia conta com a presença do ministro Ferreira do

Amaral, das Obras Públicas Transportes e Comunicações, e tem lugar às 11h00 no salão nobre dos paços do concelho. O projecto agora iniciado decorre até 1997 e prevê o realojamento de 307 agregados familiares.

2. Público - 19940505–924
   Volumes Moderados A Bolsa de Frankfurt encerrou em baixa mas dentro de um limite que os analistas consideram aceitáveis. Não se verificou pânico algum, apenas um desinteresse quase total dos investidores pelo mercado, em especial o accionista, o que explica os reduzidos montantes intermediados. O índice DAX- 30 caíu 0,15 por cento, para se estabelecer nos 2249,02 pontos. As acções mais equilibradas foram as ligadas ao sector financeiro e químico.

3. Público - 19940505–071
   Monsaraz Quer Parque Cultural A criação de um parque cultural na zona envolvente da vila de Monsaraz, que agrupe os elementos arqueológicos que venham a ser retirados da área submersa pela Barragem de Alqueva, é um dos objectivos do município de Reguengos de Monsaraz, a par da classificação da antiga povoação medieval como património rural de interesse mundial pela UNESCO. Os responsáveis da autarquia alentejana explicaram, no decurso da visita a Monsaraz do director do Centro do Património Mundial da UNESCO, Bernard Von Droste, que pretendem instalar na zona envolvente da vila histórica um parque cultural para receber os vestígios arqueológicos a retirar da área que ficará submersa pelas águas da albufeira da futura Barragem de Alqueva. Durante a visita, Von Droste revelou-se "impressionado" com a riqueza patrimonial de Monsaraz, e enalteceu o trabalho desenvolvido no âmbito da recuperação e valorização do património local. Portugal, segundo afirmou à Lusa o mesmo responsável da UNESCO, "merece um lugar de destaque na lista dos bens classificados como Património Mundial".

4. Público - 19941011–83
   Convite ao acidente As linhas do caminho-de-ferro junto à FIL, ao centro Cultural de Belém e ao longo de todo o extenso percurso da Avenida da Índia, até Algés, não têm qualquer protecção, quer de um lado quer do outro. Todos os arames estão cortados e a passagem dos carris faz-se agora com total convite ao acidente. É, no mínimo, lamentável que, quando se fala tanto em segurança no trabalho e nas estradas, não se dificulte com normais vedações a travessia da linha de comboio numa avenida tão movimentada, que serve grande parte da nossa mais importante zona turística ribeirinha, escreve-nos um leitor de Lisboa. Perante a situação, que considera de total incúria, desleixo e irresponsabilidade, o leitor pergunta ainda: será que a CP não sabe que é aquela a zona onde mais suicídios e acidentes se têm dado? Ou estará a CP à espera que alguém morra trucidado por um comboio para invocar depois razões técnicas ou vedações em estudo (há mais de um ano)? Américo Guerreiro Lisboa

5. Público - 19941012–011

Nobel da Literatura anunciado amanhã O prémio Nobel da Literatura de 1994 será anunciado amanhã, em Estocolmo, às 13h00 locais (mesma hora de Lisboa), anunciou ontem a Academia Real da Suécia, que atribui o galardão. O prémio da literatura é o único da série dos Nobel cuja data de anúncio não é fixada ao mesmo tempo que a dos outros galardões. A data é conhecida apenas com 48 horas de antecedência e é divulgada pela Academia Sueca, sendo o prémio tradicionalmente anunciada à quinta-feira. Como para os restantes prémios, o laureado receberá o montante de sete milhões de coroas suecas (cerca de 150 mil contos). No ano passado, a Academia Sueca escolheu a romancista norte-americana Toni Morrison, que se tornou a oitava mulher laureada com um prémio Nobel da Literatura desde 1901, data da criação do galardão.

6. Público - 119941025–045
   Equilíbrio no mercado O comportamento do mercado monetário foi ontem caracterizado pelo equilíbrio entre a oferta e a procura de fundos, nomeadamente no prazo dentro do qual decorre o actual período de contagem de reservas de caixa. O Banco de Portugal manteve a intenção de ceder liquidez ao sistema, em regime de leilão da taxa juro. As instituições de crédito absorveram 7,910 milhões de contos, à taxa média de nove por cento. Em comparação com a última intervenção da autoridade monetária, assistiu-se a um decréscimo do montante e à manutenção da taxa de juro. A esmagadora maioria das operações concentraram-se nos prazos até aos sete dias, com a taxa de juro a oscilar entre os nove por cento e os 9,125 por cento, enquanto nas maturidades mais dilatadas não se registou grande movimento, fixando as taxas Lisbor em 9,5940 por cento, 10,2345 por cento, 10,50 por cento e 10,6565 por cento respectivamente a um, três, seis meses e um ano. Quanto à dívida pública de curto prazo, realizou-se um leilão de Bilhetes de Tesouro a 182 dias no montante de 20 milhões de contos situando-se a taxa média nos 10,3591 por cento, reflectindo uma quebra de cerca de 1/16 pontos percentuais quando comparada com a última emissão para o mesmo prazo.

7. Público - 19950416–032
   Da CIA para os hospitais Peritos em tecnologias de informação e em satélites-espiões da Central Intelligence Agency (CIA) dos Estados Unidos estão a trabalhar juntamente com radiologistas e oncologistas norte-americanos para tentar adaptar ao diagnóstico do cancro da mama certas técnicas de tratamento de imagem que têm pertencido até aqui apenas à panóplia dos espiões. Uma das técnicas que está a ser objecto de estudo consiste num programa de computador que permite comparar duas imagens digitais. O programa tem sido usado para comparar imagens de satélites ou de aviões de reconhecimento, de forma a determinar se houve ou não movimento de tropas numa dada região, mas os mesmos princípios podem ser adaptados à comparação de mamografias (radiografias dos seios) de forma a verificar se houve ou não aparecimento de algum tumor microscópico desde o último exame. Um dia, a mesma tecnologia poderá ser

apurada de forma a determinar, apenas com base na imagem, se o tumor detectado é ou não maligno.

8. Público - 19950705–167
"BUZINÃO" NO MERCADO DE COIMBRA A inauguração do Mercado Abastecedor de Coimbra (MAC) foi ontem, ao fim da tarde, interrompida inesperadamente por um "buzinão". A cerimónia já ia adiantada quando mais de três dezenas de viaturas pesadas "irromperam" pelas ruas do mercado, buzinando incessantemente. Eram operadores do novo mercado que protestavam contra a troca de lugares de venda a que tinham sido sujeitos, contra a sua vontade. Segundo afirmam, quando em 1993 adquiriram em planta os seus postos de venda, escolheram lugares no pavilhão C, que tem melhores condições para o exercício da actividade. Mas agora, com a entrada em funcionamento do mercado, foram encaminhados para o pavilhão D, onde os espaços são mais exíguos e o acesso automóvel mais difícil. Com a manifestação de ontem, conseguiram agendar para hoje uma reunião entre os seus representantes legais e o director técnico do mercado.

9. Público - 19950912–002
Educadores no desemprego Há mais de dois mil educadores de infância no desemprego. Quem o afirma é a Fenprof num comunicado onde chama a atenção para a degradação da situação destes profissionais num sector onde, de há sete anos a esta parte, existe um congelamento na criação de novos jardins de infância. Segundo a federação, há mais de mil lugares que não vêm a concurso, porque o Ministério da Educação, alheando-se do problema, deixou a colocação dos educadores ao critério das Câmaras municipais. Estas, por sua vez, limitam-se a reconduzir nos lugares os educadores que já lá se encontravam.

10. Público - 19950924–121
O autógrafo de Narciso NA PEIXARIA do mercado de Vila do Conde, onde estiveram candidatos do PS-Porto na manhã de ontem, os bonés da "Nova Maioria"que Fernando Gomes distribuiu acabaram ao fim de poucos minutos. Como a procura era muita, alguns rapazes e raparigas da Juventude Socialista tiveram que ceder os seus bonés aos pedidos mais insistentes. O problema estava no autógrafo de Narciso Miranda na pála de alguns deles. A certa altura, uma mulher das Caxinas, que já tinha boné, não estava disposta a deixar Gomes, Narciso e Mário de Almeida sem levar um para a amiga. Só que Almeida não gostou de ler a assinatura do camarada Narciso no boné da senhora. "Ai o Narciso é que risca aqui!", comentou, azedo, o presidente da Câmara local. Na verdade, a mulher estava a borrifar-se para o autógrafo. O autarca hesitou, mas acabou por lhe arranjar um boné (pedido a outro elemento da JS). Só que também este estava autografado pelo autarca de Matosinhos. Apaziguador, Gomes lá convenceu Mário de Almeida a rubricar os dois bonés.

11. Público - 19950422–141

Acções desvalorizadas. Apesar de ter iniciado a semana passada em alta, a Bolsa de Madrid viria a encerrar este período com uma queda generalizada nas cotações. Terça-feira o mercado abriu um ciclo de perdas, que só viria a encerrar na sexta-feira quando os investidores entenderam que os descontos já tinham sido dados e as cotações voltaram a subir. Nesta sessão o índice Geral fechou nos 276,06 pontos, menos 0,15 por cento face ao último valor da semana anterior.

12. Público - 19950423–011
Ugly Kid Joe com Bon Jovi. O grupo norte-americano de rock Ugly Kid Joe junta-se à banda Van Halen para fazerem a primeira parte do concerto dos Bon Jovi, a realizar a 15 de Junho, no Estádio de Alvalade, em Lisboa. Os Ugly Kid Joe, que já actuaram em Portugal, na abertura do concerto dos Def Leppard, em Cascais, preparam-se para editar, no início de Maio, um novo álbum, de nome "Menace to Sobriety". A Tournée, que organiza a vinda dos Bon Jovi a Lisboa, prevê ainda trazer uma banda portuguesa ao Estádio de Alvalade. O objectivo é transformar a noite de 15 de Junho num "verdadeiro festival", como cita a agência Lusa.

13. Público - 19950629–083
Humberto Coelho recusa Cascais. O ex-futebolista Humberto Coelho não aceitou ocupar o lugar de vereador pelo PSD na Câmara de Cascais. Segundo explicou ao PÚBLICO o ex-jogador do Benfica, a decisão, que já foi comunicada à comissão política local dos sociais-democratas, deve-se aos muitos afazeres profissionais com a sua escola e por não ter "carreira política para estar na oposição". "Quando me candidatei era para ganhar e ter uma participação activa em termos esportivos", justificou, acrescentando "ser impossível de todo" a aceitação do cargo sem competências, devido à intensa actividade desenvolvida pela sua escola de futebol, que deverá estender a sua acção à Madeira. Perante a recusa do exfutebolista, o lugar vago de vereador "laranja" na autarquia cascaense será ocupado por Rui Libório.

14. Público - 19950629-119
Herman José continua de "Parabéns". Herman José vai continuar a apresentar o programa Parabéns durante mais um ano. O contrato com a RTP é hoje assinado no edifício da 5 de Outubro, pondo termo a todas as especulações que davam como provável a sua passagem para a SIC. Apesar de todos os convites que lhe têm sido feitos, Herman sempre insistiu que seria difícil abandonar a televisão onde sempre trabalhou. "Só poderei prescindir da RTP quando ela me começar a tratar mal, o que não tem acontecido", dizia ao PÚBLICO há cinco dias. O concurso Com a Verdade me Enganas, que era transmitido à tarde no Canal 1, acabou, mas ainda não foi desta que a televisão pública deixou escapar o "verdadeiro artista".

15. Público - 19951011–139
CLUBES PORTUGUESES MULTADOS PELA UEFA – Benfica, FC Porto e a Federação Portuguesa de Futebol (FPF) foram ontem multados pela Comissão de Controlo

e Disciplina da UEFA para os torneios europeus, que analisou os incidentes ocorridos durante os encontros realizados desde Setembro. Por conduta imprópria dos espectadores e em referência ao jogo com o Lierse, o Benfica foi multado em cinco mil francos suíços (650 contos), enquanto à FPF foi aplicada a sanção pecuniária de 390 contos (encontro com a Irlanda do Norte) e ao FC Porto, 130 contos (jogo frente ao Aalborg).

16. Público - 19951011–150
Paulo Portas, deputado independente pelo Partido Popular. É o penúltimo passo da estratégia que delineou no início do ano. Primeiro, percebeu que o PSD tinha perdido o contacto com a sociedade e, portanto, a maioria. Segundo, decidiu ilibar-se do castigo e imolar outro. Terceiro, arranjou o simpático dr. Nogueira para se queimar. Quarto, assistiu aos escombros do partido dele. O quinto é este passo, candidatar-se, esperando que o povo tenha memória curta e ficando-se pela promissão dos substitutos. Tudo isto prova que é uma candidatura profundamente viciada no passado e cujo único propósito, aliás, é arranjar uma enorme instabilidade ao país. A mim, como pessoa de direita, o candidato Cavaco Silva não me diz nada.

17. Público - 19951114–163
MIRAMAR CONDENADA NO TRIBUNAL DO TRABALHO – A Rádio Miramar foi ontem condenada pelo Tribunal do Trabalho a pagar 2091 contos de indemnizações a trabalhadores (além de mais 200 contos de multas) a quem não era reconhecido o direito a férias e respectivo subsídio, visto receberem através de recibos verdes. O Tribunal concluiu que a relação laboral entre a empresa e os trabalhadores não era a de uma simples prestação de serviços (o que libertaria a estação desses encargos), na medida em que tinham um ordenado e salário fixos, obedeciam a uma hierarquia e recebiam directivas sobre o seu trabalho.

18. Público - 19951114–169
FC PORTO CONTRATA CAMPEÃO DO MUNDO DE BILHAR– O FC Porto garantiu a contratação do campeão do mundo de bilhar, o espanhol Daniel Sanchez, que em Abril ganhou a Taça do Mundo no Porto. Sanchez, que tem 22 anos e jogava pelo Córdova, iniciou-se na modalidade aos oito anos. «Trata-se de uma carta fora do baralho, pois normalmente os jogadores só conseguem mostrar esta valia quando são mais velhos. A sua contratação foi possível com o apoio publicitário, mas o dinheiro acabou por ser indiferente na sua adesão ao clube face às propostas muito mais vantajosas que tinha de toda a Europa. Só que ele considera a nossa organização bilharística das mais bem preparadas», explicou ao PÚBLICO o vicepresidente Alípio Jorge.

19. Público - 19951220–045
Mais de dois mil contos em furtos. Ascende a mais de dois mil contos o valor global

dos nove furtos ocorridos, segunda-feira e na madrugada de ontem, em estabelecimentos e residências da zona da Grande Lisboa. Os furtos – que aconteceram maioritariamente na área da capital e apenas três nos concelhos de Loures e Cascais – não pouparam a instituição judicial. Da 10ª Vara do Tribunal da Boa Hora, em Lisboa, os assaltantes levaram material diverso, cuja lista estava ainda ontem a ser elaborada.

20. Público - 19951229–044
    Comboio mata na Moita. Um homem, de 72 anos de idade, foi, na manhã de segunda-feira, colhido mortalmente por um comboio, quando atravessava a linha na zona da estação de caminho-de-ferro da Moita, na margem sul do Tejo. Segundo a GNR, o corpo da vítima ficou feito em pedaços, que ficaram espalhados num raio de 300 metros. O homem, quando foi trucidado, levava na mão um saco com restos de comida que ia levar, como seria seu hábito, a uns cães de guarda a um armazém.

A.2 Corpus Jornal Folha de São Paulo 1994/1995 – Corpus Avaliação/Teste –

1. FSP940101-132.txt
   A PANTERA COR DE ROSA VOLTA AOS CINEMAS. "O Filho da Pantera Cor de Rosa, com direção de Blake Edwards, estréia hoje na cidade. O filme mostra as atrapalhadas aventuras do filho ilegítimo do inspetor Closeau, na investigação do rapto de uma princesa. No papel principal o ator italiano Roberto Benigni (foto). Em cartaz nos cines Gemini 1, Belas Artes e circuito.

2. FSP940101-124.txt
   DAVID SIMS. fotógrafo de moda, 26 anos. O fotógrafo inglês David Sims, hoje um dos nomes da nova geração da fotografia de moda voltada para a realidade, chegou em Londres aos 17 anos com US$ 15 (cerca de CR$ 500,00) no bolso, vindo de Sheffield, sua cidade natal. Sims nunca pensou em ser fotógrafo. Queria ser desenhista de quadrinhos. Mudou de idéia depois que viu as fotos de Larry Clark e resolveu estudar fotografia. O trabalho de Sims, assim como de Mario Sorrenti e de Corinne Day, trouxe para a fotografia um frescor até então nunca explorado. Sims, cujo trabalho inspirou estilistas como o francês Jean Colonna e o alemão Helmut lang, explica eu trabalho referindo-se à HQ: "minhas fotos são iguais aos quadrinhos, com as proporções sempre erradas. Acho legal que as pessoas hoje em dia aceitem essa estranheza". (EJ).

3. FSP940101-107.txt
   Da Reportagem Local. A brasileira Carmem de Oliveira, 28, disse após a prova feminina ter ficado feliz com o 2.° lugar. "Queria melhorar um minuto. Melhorei quase quatro", disse. Em 92, fez a prova em 54min19; ontem, levou 50min31. "Na

outra vez, ela perdeu. Hoje (ontem), a outra ganhou", disse seu técnico, Brian Appell. (MD). A edição saiu com a data do alto errada. A data correta é 1 de janeiro de 1994. ERRAMOS.

4. FSP940101-102.txt
   Da Reportagem Local. A fisiologia é o estudo das funções orgânicas de um ser vivo, neste caso o homem. O fisiologista pode ser diplomado em ciências biomédicas ou medicina. Sua formação é diferente do preparador físico, que normalmente tem diploma de educação física. O trabalho de um fisiologista tem duas etapas. A primeira é avaliar o estágio atlético de uma pessoa, atleta ou não. Para isso, são feitos os famosos testes de bicicleta e esteira, que fornecem grandezas como limiar anaeróbico (velocidade máxima que pode se obter sem que o organismo necessite de mais oxigênio do que o corpo é capaz de absorver), potência muscular (que revela o pico de desempenho muscular) e resistência muscular (a capacidade do organismo em manter-se ao longo do tempo perto do pico). Com as informações, o fisiologista define as linhas de um programa de condicionamento físico. Num clube de futebol, o programa é passado ao preparador físico, a quem cabe detalhá-lo e pô-lo em prática. A relação entre os dois profissionais é de coordenação e não de subordinação. (MD).

5. FSP940101-095.txt
   THALES DE MENEZES. Da Reportagem Local. A Copa Hopman é sempre interessante. É um torneio festivo, disputado com muita descontração. Na edição do ano passado, a Alemanha deu um passeio, vencendo fácil com a charmosa dupla Steffi Graf e Michael Stich. Neste ano, a principal atração passa a ser uma inovação tecnológica, o "juiz eletrônico"em todas as linhas da quadra. É um grande passo adiante do "sensor de solo", um sistema eletrônico mais simplório usado desde 1987 nas quadras centrais de alguns templos do tênis mundial, mas restrito às linhas de serviço o que ajuda muito em saques a 200 km/h. O novo sistema tem dois desafios. Um deles, exclusivamente técnico, é demonstrar eficiência nos testes de campo. O outro desafio é mais complicado: a desconfiança de jogadores e público. É pena que John McEnroe desistiu de participar do torneio. Seria o piloto-detestes perfeito para o novo sistema. "Big Mac"foi o primeiro a testar o "sensor de solo"em 1987, proporcionando uma das cenas mais hilariantes do esporte. Revoltado, ficou xingando as linhas da área de saque como se fossem pessoas, enquanto a platéia estourava de tanto rir. Palhaçadas à parte, a cena protagonizada por McEnroe é simbólica. Quem acompanha os bastidores do circuito sabe que todos os jogadores odeiam o sensor. É lógico, já que um juiz humano pode ser intimidado por reclamações. E o público? Numa pesquisa no US Open 91, 75% dos entrevistados votou contra o sensor. Afinal, todo mundo gosta de vaiar numa bola duvidosa. Para os tradicionalistas, só resta torcer para muitas falhas no sistema.

6. FSP940101-092.txt
   Da Sucursal do Rio. O inquérito policial para investigar as denúncias de corrupção

no futebol do Rio foi instaurado a pedido do deputado estadual Sérgio Cabral Filho (PSDB). "Estou esperançoso, mas só lamento que a Assembléia Legislativa não tenha entrado nesta apuração", disse o autor do projeto de criação da chamada "CPI do apito". Com a instauração do inquérito pela polícia do Rio, as investigações de supostas irregularidades no futebol fluminense se ampliam. Também a pedido de Cabral Filho, o Ministério Público do Estado está investigando as denúncias envolvendo o diretor da Comissão de Arbitros, Wagner Canazaro, e o presidente da Federação, Eduardo Vianna. De acordo com as denúncias dos árbitros, o esquema que seria coordenado por Canazaro e Vianna favoreceria clubes ligados aos dois. A criação da "CPI do Apito"foi adiada depois que dois deputados ligados a Vianna, Almir Rangel (PSC) e Albano Reis (sem partido), apresentaram emendas ao projeto. Dependendo do calendário, a decisão ficará para fevereiro, após o recesso parlamentar.

7. FSP940101-085.txt
   Da Reportagem Local. Para este ano a maior novidade no setor dos transportes no município deve ser a introdução do sistema de catracas eletrônicas, que poderá gerar demissões nas empresas de transporte, já que os cobradores não serão mais necessários. O sindicato dos condutores é contra a medida e, no ano passado, ameaçou fazer greves em protesto. Com a catraca eletrônica, a compra de bilhetes poderia ser antecipada, como no Metrô, e também permitiria integração gratuita entre uma linha e outra. O programa de corredores, outra promessa para este ano, teve as primeiras licitações lançadas no final do ano passado. Vai permitir a integração mais rápida dos bairros através da circulação dos ônibus em faixas exclusivas. Se tudo der certo, os corredores devem começar a ser implantados no final do ano.

8. FSP940101-079.txt
   Da Redação. A pesquisa Datafolha é um levantamento por amostragem estratificada, com cotas de sexo e cidade, sendo que o conjunto da população adulta da cidade é tomado como universo da pesquisa. Neste levantamento, realizado entre os dias 16 e 17 de dezembro, foram entrevistadas 1.076 pessoas em São Paulo, 646 no Rio de Janeiro, 432 em Belo Horizonte, 432 em Salvador, 432 em Porto Alegre, 432 em Curitiba, 432 em Florianópolis, 432 em Fortaleza, 428 em Recife e 431 em Campo Grande. A direção do Datafolha é exercida pelos sociólogos Antonio Manuel Teixeira Mendes e Gustavo Venturi, tendo como assistentes Mauro Francisco Paulino, Emilia de Franco e a estatística Renata Nunes Cesar.

9. FSP940101-074.txt
   Da Reportagem Local. O shopping Center Norte vai sortear uma viagem ao Caribe. Para concorrer, é preciso trocar notas fiscais recebidas durante as compras em lojas do shopping por cupons. Cada CR$ 5.000,00 em notas vale um cupom, que ficará depositado na urna do shopping até as 18h do dia 30 de janeiro, quando acontece o sorteio. O prêmio é um cruzeiro pelo Caribe, com direito a um acompanhante e todas

as despesas pagas. O slogan da promoção é "Que tal catar coquinho no Caribe?". Essa é a última viagem sorteada pelo shopping. Todos os meses, desde junho, o Center Norte dá como prêmio uma viagem internacional a seus frequentadores. A promoção vale para as compras feitas a partir de segunda-feira. Os cupons serão trocados no posto do shopping apenas até o dia 29, véspera do sorteio.

10. FSP940101-066.txt
Da Folha Nordeste. A Fundação Civil Casa de Misericórdia, que administra a Santa Casa de Franca, vai assumir a área da Saúde do município por 30 dias, até que a prefeitura local resolva a situação do setor. A decisão foi anunciada ontem após reunião entre representantes da fundação e o prefeito Ary Balieiro. "Esses 30 dias serão suficientes para reassumirmos a Saúde do município", afirmou o prefeito. A crise do setor em Franca agravou-se anteontem com a demissão do secretário Ciro Botto. "Não aceitei ainda a decisão do Botto. Ele continua no cargo até segunda ordem", disse ontem Balieiro, do PMDB. A greve dos médicos municipais, há dois meses, iniciou a crise da Saúde em Franca. O problema agravou-se na última sexta-feira, com a demissão coletiva dos grevistas.

A.2.1 Jornal Folha de São Paulo–Textos/1995 – Corpus Avaliação/Teste –

1. FSP950101-011.txt

A Folha, em editorial na quarta-feira sob o título "Chega de roubalheira", comenta a apresentação do relatório final da Comissão Especial de Investigação (CEI) sobre suspeitas de irregularidades no Executivo. O editorial afirma que é razoável imaginar que o relatório não revele mais que a "ponta do iceberg", mas que oferece um mapa inicial da corrupção na administração pública. "No caso dos transportes, por exemplo, um dos setores mais visados no relatório, o sobrepreço médio na construção de estradas, segundo a CEI, é de 40próximo governo será então posto à prova desde o seu início, com o desafio de mostrar, com celeridade e ações concretas, se vai ou não compactuar com o binômio corrupção-impunidade que há tanto sangra o país".

2. FSP950101-032.txt

O cenário de curto prazo indica a manutenção do clima de turbulência nas Bolsas de Valores, com as intervenções do Banco Central no Banespa e no Banerj. Elas se somam ao impacto negativo da crise mexicana sobre os investimentos estrangeiros nas bolsas latinoamericanas. Na última semana do ano, o Índice Bovespa registrou alta de apenas 0,06%, mas acumula desvalorização de 6,49% em dezembro e de 20,60%

nos últimos três meses. O Índice Senn, da Bolsa de Valores do Rio, subiu 2,73% na semana mas fechou o mês com queda de 3,50%.

3. FSP950101-054.txt
   ZÉLIA GATTAI, 78, escritora - "É muito difícil escolher. Em 45, conheci Jorge, meu primeiro filho nasceu em 41, os outros dois filhos nasceram em 47 e 51. Todos estes anos foram maravilhosos. O melhor réveillon foi em 52, quando passamos no Kremlin. Havia escritores, artistas e músicos de toda parte do mundo no palácio. O baile nos grandes salões me impressionou muito.".

4. FSP950101-084.txt

   20 de dezembro de 1976. Prezado Senhor Caro,. Muito obrigado pela carta. O que o senhor escreve sobre problemas de tradução evidentemente me interessa muito. Infelizmente nada sei de português, um idioma que mal conheço. Apesar disso, desejo, por ocasião de sua tradução do "Auto da Fé", iniciar-me nessa língua. Hoje, na verdade, escrevo para comunicar-lhe o número do meu telefone em Zurique: 47-0936. Certamente passarei aqui a maior parte de janeiro, em função de uma doença séria de minha mulher (ela mora e trabalha sempre em Zurique). Por favor, telefone, e caso minha mulher já esteja melhor, como espero que aconteça, terei enorme satisfação em conhecê-lo. Meus melhores cumprimentos,. Seu, Elias Canetti. Traduções de ANDRÉ CARONE.

5. FSP950111-014.txt

   Cerca de 200 policiais procuram no norte de Minas Gerais os fazendeiros Darly Alves e seu filho Darci Alves Pereira. Os dois foram condenados a 19 anos de prisão cada um, em dezembro de 90, pelo assassinato do líder seringueiro Chico Mendes. O crime ocorreu em dezembro de 88 no município de Xapuri (AC). A polícia de Minas iniciou as buscas em dezembro, após uma denúncia anônima. Segundo informações recebidas pela polícia, os dois estariam escondidos em uma fazenda de difícil acesso.

6. FSP950111-026.txt

   Da Sucursal do Rio. A juíza Marilene Soares Reis Franco, da 13ª Vara Federal, acatou denúncia por formação de quadrilha e estelionato contra os secretários de Planejamento do Rio, Marco Aurélio Alencar, e da Fazenda, Edgar Gonçalves da Rocha. Marco Aurélio é filho do governador Marcello Alencar (PSDB). Os dois foram denunciados pelo Ministério Público, acusados de irregularidades na aplicação de recursos da Prefeitura do Rio entre outubro de 90 e setembro de 91 (na gestão Marcello Alencar). Na época, Marco Aurélio foi assessor especial de Alencar na área financeira e

Edgar da Rocha, secretário municipal de Fazenda. A denúncia foi feita em 6 de dezembro pelo procurador da República Alex Miranda. O governador é uma das testemunhas arroladas por Miranda para serem ouvidas pela Justiça Federal. (Francisco Santos e Aziz Filho).

7. FSP950111-034.txt

Da Sucursal de Brasília . Pressões políticas fizeram o governo adiar ontem, por tempo indeterminado, a posse dos presidentes do Banco do Brasil, Paulo César Ximenes, e da Caixa Econômica Federal, Sérgio Cutolo. A posse de Ximenes estava marcada para hoje, e a de Cutolo, para amanhã. Aliados do governo vêm reivindicando cargos nos bancos. A equipe econômica quer que todas as diretorias sejam ocupadas por técnicos. Por isso, o ministro da Fazenda, Pedro Malan, decidiu só empossar Ximenes e Cutolo quando todos os diretores estiverem escolhidos. A equipe econômica decidiu apressar a escolha das diretorias. Ximenes e Cutolo já fizeram as indicações. Os nomes podem ser aprovados até a próxima semana.

8. FSP950111-036.txt

O novo presidente da Câmara dos Deputados dos EUA, Newt Gingrich, demitiu na noite de segunda-feira a historiadora da Câmara, Christina Jeffrey, recém-indicada por ele. O motivo seria seu apoio a um bloqueio de fundos federais para o estudo do Holocausto nas escolas por não levar em conta a visão dos nazistas e dos membros da Ku Klux Klan.

9. FSP950117-048.txt

O custo médio dos 68 produtos que fazem parte da cesta básica do paulistano ficou ontem em R$ 99,02, com alta de 0,12% sobre os R$ 98,90 de sexta-feira. No mês, a cesta acumula queda de 3,60pesquisa é realizada pelo Procon, em convênio com o Dieese. O grupo alimentação ficou estável ontem, apesar de as maiores altas terem sido de produtos que compõem o grupo, que no mês acumula redução de 4,46%. Os grupos de higiene pessoal e de produtos de limpeza apresentaram aumentos ontem de 1,26% e 0,22%, respectivamente. Os produtos que mais subiram ontem foram salsicha avulsa (6,54%), biscoito de maisena Triunfo (4,55%) e frango resfriado inteiro (1,59%). As maiores quedas ontem foram da carne de primeira sem osso/acém (- 2,51%), leite em pó integral Itambé (-1,83%) e carne de primeira/coxão mole (-0,90%). Na pesquisa de ontem do Procon/Dieese em 70 supermercados da cidade de São Paulo, 26 produtos subiram, 15 baixaram de preço e 27 permaneceram estáveis. O custo mínimo da cesta ontem foi de R$ 67,99, enquanto o custo máximo chegou aos 144,71.

10. FSP950117-074.txt

    Free-lance para a Folha. Maria da Glória Chagas Pereira, 32, foi presa em flagrante no início da madrugada de ontem, no Rio, ao tentar matar a facadas sua mãe, Ana Maria Chagas Pereira, 60. A tentativa de homicídio ocorreu em uma casa no bairro do Realengo (zona norte do Rio). Esfaqueada na barriga, Ana Maria foi socorrida por vizinhos, que chamaram a polícia e a levaram para o hospital Albert Schweitzer, no mesmo bairro. Ontem, seu estado de saúde era regular. Na 33ª DP (Realengo), Maria da Glória contou que sofreu uma crise nervosa e por isso atacou sua mãe. Segundo policiais, ela sofre de problemas mentais e teria se descontrolado depois de ingerir vários medicamentos. (Ronaldo Soares).

Apêndice B

Relações Retóricas - RST

RST - Teoria da Estrutura Retórica
Mann and Thompson–1987–1988
e Taboada and Mann–2005–2006

Definições das relações de apresentação

Nome da relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Antítese	em N: A tem atitude positiva face a N	N e S estão em contraste (cf. a relação de Contraste); devido à incompatibilidade suscitada pelo contraste, não é possível ter uma atitude positiva perante ambas as situações; a inclusão de S e da incompatibilidade entre as situações aumenta a atitude positiva de L por N	A atitude positiva do L face a N aumenta

Continua na próxima página...

Nome da relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Concessão	em N: A possui atitude positiva face a N em S: A não afirma que S não está certo	A reconhece uma potencial ou aparente incompatibilidade entre N e S; reconhecer a compatibilidade entre N e S aumenta a atitude positiva de L face a N	A atitude positiva de L face a N aumenta
Elaboração	em N: apresenta uma acção de L (incluindo a aceitação de uma oferta), não realizada face ao contexto de N	A compreensão de S por L aumenta a capacidade potencial de L para executar a ação em N	A potencial capacidade de L para executar a acção em N aumenta
Evidência	em N: L pode não acreditar em N a um nível considerado por A como sendo satisfatório em S: L acredita em S ou considera-o credível	A compreensão de S por L aumenta a crença de L em N	A crença de L em N aumenta
Fundo	em N: L não compreende integralmente N antes de ler o texto de S	S aumenta a capacidade de L compreender um elemento em N	A capacidade de L para compreender N aumenta
Justificação	nenhuma	A compreensão de S por L aumenta a sua tendência para aceitar que A apresente N	A tendência de L para aceitar o direito de A a apresentar N aumenta
Motivação	em N: N é uma ação em que L é o ator (incluindo a aceitação de uma oferta), não realizada face ao contexto de N	A compreensão de S aumenta a vontade de L para executar a ação em N	A vontade de L para executar a ação em N aumenta

Continua na próxima página...

Nome da relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Preparação	nenhuma	S precede N no texto; S tende a fazer com que L esteja mais preparado, interessado ou orientado para ler N	L está mais preparado, interessado ou orientado para ler N
Reformulação	nenhuma	em N + S: S reformula N, onde S e N possuem um peso semelhante; N é mais central para alcançar os objetivos de A do que S	L reconhece S como reformulação
Resumo	em N: N deve ser mais do que uma unidade	S apresenta uma reformulação do conteúdo de N, com um peso inferior	L reconhece S como uma reformulação mais abreviada de N

Tabela B.1: Relações retóricas propostas na RST

Definições das relações de conteúdo

Nome da relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Alternativa (anti-condicional)	em N: N representa uma situação não realizada em S: S representa uma situação não realizada	realização de N impede a realização de S	L reconhece a relação de dependência de impedimento que se estabelece entre a realização de N e a realização de S
Avaliação	nenhuma	em N + S: S relaciona N com um grau de atitude positiva de A face a N	L reconhece que S confirma N e reconhece o valor que lhe foi atribuído

Continua na próxima página...

Nome da relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Causa involuntária	em N: N não representa uma acção voluntária	S, por outras razões que não uma ação voluntária, deu origem a N; sem a apresentação de S, L poderia não conseguir determinar a causa específica da situação; a apresentação de N é mais importante para cumprir os objetivos de A, ao criar a combinação N-S, do que a apresentação de S	L reconhece S como causa de N
Causa voluntária	em N: N constitui uma acção voluntária ou mesmo uma situação possivelmente resultante de uma acção voluntária	S poderia ter levado o agente da acção voluntária em N a realizar essa acção; sem a apresentação de S, L poderia não perceber que a ação fui suscitada por razões específicas ou mesmo quais foram essas razões; N é mais importante do que S para cumprir os objetivos de A, na criação da combinação N-S	L reconhece S como a causa da ação voluntária em N
Circunstância	em S: S não se encontra não realizado	S define um contexto no assunto, no âmbito do qual se pressupõe que L interprete N	L reconhece que S fornece o contexto para interpretar N
Condição	em S: S apresenta uma situação hipotética, futura, ou não realizada (relativamente ao contexto situacional de S)	Realização de N depende da realização de S	L reconhece de que forma a realização de N depende da realização de S

Continua na próxima página...

Nome da relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Condição inversa	nenhuma	S afecta a realização de N; N realiza-se desde que S não se realize	L reconhece que N se realiza desde que S não se realize
Elaboração	nenhuma	S apresenta dados adicionais sobre a situação ou alguns elementos do assunto apresentados em N ou passíveis de serem inferidos de N, de uma ou várias formas, conforme descrito abaixo. Nesta lista, se N apresentar o primeiro membro de qualquer par, então S inclui o segundo: conjunto :: membro abstracção :: exemplo todo :: parte processo :: passo objecto :: atributo generalização :: especificação	L reconhece que S proporciona informações adicionais a N. L identifica o elemento do conteúdo relativamente ao qual se fornece pormenores
Incondicional	em S: S poderia afetar a realização de N	N não depende de S	L reconhece que N não depende de S
Interpretação	nenhum	em N + S: S relaciona N com várias ideias que não se encontram directamente relacionadas com N, e que não estão relacionadas com a atitude positiva de A	L reconhece que S relaciona N com várias idéias que não se encontram relacionadas com o conhecimento apresentado em N
Método	em N: uma actividade	S apresenta um método ou instrumento que tende a aumentar as probabilidades de realização de N	L reconhece que o método ou instrumento de S tende a aumentar as probabilidades de realização de N

Continua na próxima página...

Nome da relação	Condições em S ou N, individualmente	Condições em N + S	Intenção do A
Propósito	em N: N é uma atividade; em S: S é uma situação que não se encontra realizada	S será realizado através da atividade de N	L reconhece que a atividade em N se inicia para realizar S
Resultado involuntário	em S: S não representa uma ação voluntária	N causou S; a apresentação de N é mais importante para cumprir os objectivos de A, ao criar a combinação N-S, do que a apresentação de S	L reconhece que N poderia ter causado a situação em S
Resultado voluntário	em S: S constitui uma situação ou ação voluntária possivelmente resultante de uma ação voluntária	N pode ter causado S; a apresentação de N é mais importante para cumprir os objetivos de A do que a apresentação de S	L reconhece que N pode ser uma causa da ação ou situação em S
Solução	em S: S apresenta um problema	N constitui uma solução para o problema apresentado em S	L reconhece N como uma solução para o problema apresentado em S

Tabela B.2: Relações retóricas propostas na RST

Definições das relações multi-nucleares

Nome da relação	Condições em cada par de N	Intenção de A
Conjunção	Os elementos unem-se para formar uma unidade onde cada um dos elementos desempenha um papel semelhante	L reconhece que os elementos inter-relacionados se encontram em conjunto

Continua na próxima página...

Nome da relação	Condições em cada par de N	Intenção de A
Contraste	Nunca mais de dois núcleos; as situações nestes dois núcleos são (a) compreendidas como sendo as mesmas em vários aspectos (b) compreendidas como sendo diferentes em alguns aspectos, e (c) comparadas em termos de uma ou mais destas diferenças	L reconhece a possibilidade de comparação e a(s) diferença(s) suscitadas pela comparação realizada
Disjunção	Um dos elementos apresenta uma alternativa (não necessariamente exclusiva) a(s) outra(s)	L reconhece que os elementos inter-relacionados constituem alternativas
Junção	nenhuma	nenhuma
Lista	Um elemento comparável a outros e ligado a outro N através de uma relação de Lista	L reconhece a possibilidade de comparação dos elementos relacionados
Reformulação multi-nuclear	Um elemento constitui, em primeiro lugar, a repetição de outro, com o qual se encontra relacionado; os elementos são de importância semelhante aos objectivos de A	L reconhece a repetição através dos elementos relacionados
Sequência	Existe uma relação de sucessão entre as situações apresentadas nos núcleos	L reconhece as relações de sucessão entre os núcleos

Tabela B.3: Relações retóricas propostas na RST

Apêndice C

Relações Rétoricas – Daniel Marcu/2001

Attribution	attribution, attribution-negative
Background	background, circumstance
Cause	cause, result, consequence
Comparison	comparison, preference, analogy, proportion
Condition	condition, hypothetical, contingency, otherwise
Contrast	contrast, concession, antithesis
Elaboration	elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition
Enablement	purpose, enablement
Evaluation	evaluation, interpretation, conclusion, comment
Explanation	evidence, explanation-argumentative, reason
Joint	list, disjunction
Manner-Means	manner, means
Topic-Comment	problem-solution, question-answer statement-response, topic-comment,comment-topic, rhetorical-question
Summary	summary, restatement
Temporal	temporal-before, temporal-after, temporal-same-time, sequence, invert-edsequence
Topic Change	topic-shif; topic-drift

Tabela C.1:

Mononuclear (satellite)	Mononuclear (nucleus)	Multinuclear
analogy		analogy
antithesis		Contrast
attribution		
attribution-n		
background		

	cause	Cause-Result
circumstance		
comparison		Comparison
comment		
		Comment-Topic
concession		
conclusion		Conclusion
condition		
consequence-s	consequence-n	Consequence
contingency		Contrast (see antithesis)
definition		
		Disjunction
elaboration-additional		
elaboration-set-member		
elaboration-part-whole		
elaboration-process-step		
elaboration-object-attribute		
elaboration-general-specific		
enablement		
evaluation-s	evaluation-n	Evaluation
evidence		
example		
explanation-argumentative		
hypothetical		
interpretation-s	interpretation-n	Interpretation
		Inverted-Sequence
		List
manner		
means		
otherwise		Otherwise
preference		
problem-solution-s	problem-solution-n	Problem-Solution
		Proportion
purpose		
question-answer-s	question-answer-n	Question-Answer

reason		Reason
restatement		
	result	Cause-Result
rhetorical-question		
		Same-Unit
		Sequence
statement-response-s	statement-response-n	Statement-Response
summary-s	summary-n	
	temporal-before	
temporal-same-time	temporal-same-time	Temporal-Same-Time
	temporal-after	
		TextualOrganization
		Topic-Comment
topic-drift		Topic-Drift
topic-shift		Topic-Shift

Apêndice D

Macroestrutura/Macroproposição – Sistema AuTema-Dis –

Textos Jornal Público 1994–1995
Português Europeu (PE)
Corpus Aprendizado/Treino–10 textos

1. Publico-19940101-007

   Dominic Sonny Constanzo morreu. Ao longo de a sua carreira tocou com o clarinetista Woody Herman, o trompetista Thad Jones, o baterista Mel Lewis e o cantor Clark Terry. Acompanhou a vocalista Rosemary Clooney. Conseguiu fazer a primeira gravação a Stash.

2. Publico-19941012-035 Os operadores detectaram, o que se reflectiu em o fecho. Com efeito, o FTSE, recuperou 40,7 pontos.

3. Publico-19941214-076

   Escassas dezenas de estudantes participaram em uma manifestação estudantil destinada a protestar contra a detenção de dois alunos há cinco dias. Um de os jovens, havia sido detido por a PSP de Almada. Um colega acabou por ser obrigado a lá ficar também toda a noite. Alegada injúria e desobediência a os agentes estiveram em a base de as detenções. O protesto de ontem, que foi organizada por os dirigentes estudantis de a escola secundária de o Monte da Caparica, acabou por não passar de um reduzido grupo de jovens reunidos em a Praça São João Baptista. Não houve aderência, lamentava um aluno a manifestação.

4. Publico-19950519-057

   Por a Universidade de Kuopio conclui que é possível detectar a doença de Alzheimer cinco a dez anos. Kalevi Pyorala afirma que os médicos em geral diagnosticam a

doença. mas 35 a 37 por cento de os casos distinguir as perturbações de memória típicas de a doença de as provocadas por outras causas. O estudo de Pyorala, foi publicado em a revista Neurology.

5. Publico-19950725-025

   Disseram os operadores. O índice de a Bolsa de Sidney, perdeu 2,8 pontos e fechou em os 2108,7 pontos.

6. Publico-19950726-079

   O abastecimento público de água em o concelho de Elvas degradou-se ultimamente. A seca prolongada, as temperaturas de os últimos dias e o aumento de o consumo em esta época de o ano foram as razões indicadas para justificar o agravamento de as condições de o abastecimento público. A câmara nomeou, entretanto, um grupo de trabalho.

7. Publico-19950814-011

   O cineasta francês Marcel Moussy, argumentista de os filmes Os 400 Golpes, de François Truffaut e Paris está a arder? de René Clément, morreu em Caen, França, com 71 anos de idade, de doença prolongada, divulgou ontem a sua família. Marcel Moussy realizou várias obras de ficção. É também autor de várias peças de teatro e romances A sua última obra de ficção, publicada em 1990, foi galardoada com o Prémio Albert Camus. Marcel Moussy

8. Publico-19950916-121

   Recorde adiado. Foi em esta semana que caiu o recorde absoluto de a Bolsa de Londres. A maior subida registou-se As expectativas de uma descida em as taxas de juro britânicas e norte-americanas, juntamente sustentaram esta subida. Os dealers esperam que o mercado consolide.

9. Publico-19950916-157

   O castanho e preto serão as cores mais usadas em este Outono/Inverno. Botas e botins práticos, mocassins e sapatos baixos em uma linha bastante simples são as tendências para o dia-a-dia. Os sapatos clássicos têm formas arredondadas ou bicudas, muitas tiras que apertam o tornozelo e outras que cruzam em o peito do pé. Os saltos são altos tornando-se o complemento de toda uma elegância. Dominam as vitelas italianas, nobucks e crutes acamurçados. Estas são as propostas de a Hera.

10. Publico-19950917-041

    Autarcas portugueses em o combate a a cólera. A Associação Nacional de Municípios está a lançar uma campanha de solidariedade de o poder local português. O objectivo é ajudar a combater a epidemia de cólera que há vários meses assola Cabo Verde. De acordo com um comunicado de a Associação Nacional de Municípios, vai ser enviada

de imediato uma primeira remessa de medicamentos e materiais de desinfecção e de prevenção. A iniciativa tem vindo a encontrar a melhor receptividade em os municípios portugueses, refere a Associação Nacional de Municípios.

Textos Jornal Público–1994/1995
Português Europeu (PE)
Corpus Avaliação/Teste–20 Textos (PE)

1. Publico-19940504-070 A Câmara Municipal Do Montijo assina INH e IGAPHE, o acordo de adesão Per, um pacote governamental. A cerimónia conta com a presença de o ministro Ferreira do Amaral, de as Obras Públicas Transportes e Comunicações, e tem lugar. O projecto iniciado decorre até 1997 e prevê o realojamento de 307 agregados familiares.

2. Publico-19940505-024 A Bolsa de Frankfurt encerrou. Não se verificou pânico algum, o accionista, O índice DAX-30 caiu 0,15 por cento, As acções mais equilibradas foram as ligadas a o sector financeiro e químico.

3. Publico-19940505-071 A criação de um parque cultural em a zona envolvente de a vila de Monsaraz, é um de os objectivos de o município de Reguengos de Monsaraz. Os responsáveis de a autarquia alentejana explicaram. Durante a visita, Von Droste revelou-se impressionado com a riqueza patrimonial de Monsaraz, e enalteceu o trabalho desenvolvido Portugal, merece um lugar de destaque em a lista de os bens classificados como Património Mundial.

4. Publico-19941011-083

   As linhas de o caminho-de-ferro não têm qualquer protecção, quer de um lado quer de o outro. Todos os arames estão cortados e a passagem de os carris faz-se. É, escrevenos um leitor de Lisboa. O leitor pergunta será que a CP não sabe que é aquela a zona onde mais suicídios e acidentes se têm dado Ou estará a CP a a espera que alguém morra trucidado por um comboio para invocar depois razões técnicas ou vedações em estudo há mais de um ano. Américo Guerreiro Lisboa

5. Publico-19941012-011

   O prémio Nobel da Literatura de 1994 será anunciado anunciou a Academia Real da Suécia, O prémio de a literatura é o único de a série de os Nobel cuja data de anúncio não é fixada ao mesmo tempo que a de os outros galardões. A data é conhecida apenas com 48 horas de antecedência e é divulgada por a Academia Sueca, o laureado receberá o montante de sete milhões de coroas suecas cerca de 150 mil contos. A Academia Sueca escolheu a romancista norte-americana Toni Morrison.

6. Publico-19941025-045

   O comportamento de o mercado monetário foi caracterizado por o equilíbrio entre a oferta e a procura de fundos, nomeadamente dentro de o decorre o actual período de contagem de reservas de caixa. O Banco de Portugal manteve a intenção de ceder liquidez a o sistema, em regime de leilão de a taxa juro. As instituições de crédito absorveram 7,910 milhões de contos. Em comparação assistiu-se a um decréscimo de o montante e a a manutenção de a taxa de juro. A esmagadora maioria de as operações concentraram-se realizou-se um leilão de Bilhetes de Tesouro.

7. Publico-19950416-032 Peritos em tecnologias de informação e em satélites-espiões de a Central Intelligence Agency CIA de os Estados Unidos estão a trabalhar juntamente com radiologistas e oncologistas norte-americanos para tentar adaptar a o diagnóstico de o cancro de a mama certas técnicas de tratamento de imagem. Uma de as técnicas que está a ser objecto de estudo consiste em um programa de computador que permite comparar duas imagens digitais. O programa tem sido usado para comparar imagens de satélites ou de aviões de reconhecimento, ou não movimento de tropas podem ser adaptados a a comparação de mamografias radiografias de os seios a mesma tecnologia poderá ser apurada de forma a determinar, apenas com base em a imagem.

8. Publico-19950705-167

   A inauguração de o Mercado Abastecedor de Coimbra MAC foi ao fim da tarde, interrompida inesperadamente por um buzinão. A cerimónia ia adiantada quando mais de três dezenas de viaturas pesadas irromperam por as ruas de o mercado. Eram operadores de o novo mercado que protestavam contra a troca de lugares de venda a Segundo afirmam, escolheram lugares C, que tem melhores condições. Com a manifestação de ontem, conseguiram agendar para hoje uma reunião entre os seus representantes legais e o director técnico de o mercado. Automóvel mais difícil.

9. Publico-19950912-022 Há mais de dois mil educadores de infância. Quem o afirma é a Fenprof em um comunicado. Segundo a federação, há mais de mil lugares porque o Ministério da Educação, deixou a colocação de os educadores a o critério de as Câmaras municipais. Estas, por sua vez, limitam-se a reconduzir em os lugares os educadores.

10. Publico-19950924-121 Os bonés de a Nova Maioria que Fernando Gomes distribuiu acabaram a o fim de poucos minutos. Tiveram que ceder os seus bonés a os pedidos mais insistentes. O problema estava em o autógrafo de Narciso Miranda não estava disposta a deixar Gomes, Narciso e Mário de Almeida sem levar um para a amiga. Só que Almeida não gostou de ler a assinatura de o camarada Narciso em o boné de a senhora. Ai o Narciso é que risca aqui! Comentou, azedo, o presidente de a Câmara local. A mulher estava a borrifar-se O autarca hesitou, Só que também este estava

autografado por o autarca de Matosinhos. Apaziguador, Gomes convenceu Mário de Almeida a rubricar os dois bonés.

11. Publico-19950422-141

    Acções desvalorizadas. A Bolsa de Madrid viria a encerrar este período em as cotações. O mercado abriu um ciclo de perdas, e as cotações voltaram a subir o índice Geral fechou em os 276,06 pontos.

12. Publico-19950423-011

    Ugly Kid Joe com Bom Jovi. Os Ugly Kid Joe, Cascais, preparam-se para editar, em o início de Maio, de nome Menace to Sobriety. A Tournée, prevê trazer uma banda portuguesa. O objectivo é transformar a noite de 15 de Junho.

13. Publico-19950629-083 Humberto Coelho recusa Cascais. O ex-futebolista Humberto Coelho não aceitou ocupar o lugar de vereador. Segundo explicou a o Público o ex-jogador de o Benfica, a decisão, deve-se a os muitos afazeres profissionais era para ganhar e ter uma participação activa em termos desportivos, justificou, o lugar vago de vereador laranja em a autarquia cascaense será ocupado por Rui Libório.

14. Publico-19950629-119 Herman José continua de Parabéns. Herman José vai continuar a apresentar o programa Parabéns. O contrato com a RTP é assinado em o edifício de todas as especulações. Apesar de todos os convites Herman insistiu que seria difícil abandonar a televisão trabalhou. Só poderei prescindir de a RTP dizia a o Público. O concurso Com a Verdade me Enganas, acabou, deixou escapar o verdadeiro artista.

15. Publico-19951011-139 Clubes Portugueses Multados Pela UEFA. Benfica FC Porto e a Federação Portuguesa de Futebol FPF foram multados por a Comissão de Controlo e Disciplina de a UEFA, o Benfica foi multado em cinco mil francos suíços.

16. Publico-19951011-150

    É o penúltimo passo de a estratégia que delineou em o início de o ano. Primeiro, percebeu que o PSD tinha perdido o contacto com a sociedade e portanto, a maioria. Segundo, decidiu ilibar-se de o castigo e imolar outro. Terceiro, arranjou o simpático dr. Nogueira. Quarto, assistiu a os escombros de o partido de ele. O quinto é este passo, candidatar-se. Tudo isto prova que é uma candidatura profundamente viciada em o passado e cujo único propósito, aliás, é arranjar uma enorme instabilidade a o país. O candidato Cavaco Silva não me diz nada.

17. Publico-19951114-163 Miramar Condenada em o Tribunal de o Trabalho. A Rádio Miramar foi condenada por o Tribunal do Trabalho visto receberem. O Tribunal concluiu que a relação laboral entre a empresa e os trabalhadores não era a de uma simples prestação de serviços na medida em que tinham um ordenado e salário fixos, obedeciam a uma hierarquia e recebiam directivas.

18. Publico-19951114-169 FC Porto Contrata Campeão de o Mundo De Bilhar. O FC Porto garantiu a contratação de o campeão de o mundo de bilhar, Sanchez, jogava por o Córdova, iniciou-se em a modalidade. Trata-se de uma carta pois normalmente os jogadores só conseguem mostrar esta valia. A sua contratação foi possível com o apoio publicitário, tinha de toda a Europa. Só que ele considera a nossa organização bilharística de as mais bem preparadas, explicou a o Público o vice-presidente Alípio Jorge.

19. Publico-19951220-045 Ascende a mais de dois mil contos o valor global de os nove furtos ocorridos, segunda-feira e em a madrugada de em estabelecimentos e residências de a zona de a Grande Lisboa. Os furtos não pouparam a instituição judicial. Vara do Tribunal da Boa Hora, os assaltantes levaram material diverso, em furtos.

20. Publico-19951229-044 Comboio mata em a Moita. Um homem, foi, colhido mortalmente por um comboio, o corpo de a vítima ficou feito em pedaços. O homem, levava em a mão um saco.

Textos Jornal Folha de São Paulo–1994/1995
Português Brasileiro (PB)
Corpus Avaliação/Teste–20 textos

1. FSP950101-011 A Folha, comenta a apresentação de o relatório final de a Comissão Especial de Investigação CEI. O editorial afirma que é razoável imaginar que o relatório não revele que a ponta de o iceberg, diz o editorial. O próximo governo será posto a a prova não compactuar com o binômio corrupção-impunidade.

2. FSP950101-032 O cenário de curto prazo indica a manutenção de o clima de turbulência. O Índice Bovespa registrou alta de apenas 0,06 , O Índice Senn subiu 2,73 latino-americanas.

3. FSP950101-054 ZÉLIA GATTAI, 78, escritora. É muito difícil escolher conheci Jorge, meu primeiro filho nasceu em 41, os outros dois filhos nasceram em 47 e 51. Todos estes anos foram maravilhosos. O melhor réveillon foi em 52, Havia escritores, artistas e músicos de toda parte de o mundo O baile em os grandes salões me impressionou muito.

4. FSP950101-084

   20 de dezembro de 1976. O que o senhor escreve sobre problemas de tradução evidentemente me interessa muito. Infelizmente nada sei de português, Apesar disso, desejo, escrevo para comunicar-lhe o número de o meu telefone em Zurique: Por favor, telefone, e caso minha mulher esteja melhor que aconteça, terei enorme satisfação Meus melhores cumprimentos. Traduções de ANDRÉ CARONE. Prezado Senhor Caro por a carta. Ela mora e trabalha em Zurique. Seu Elias Canetti.

5. FSP950111-014 Cerca de 200 policiais procuram em o norte de Minas Gerais os fazendeiros Darly Alves e seu filho Darci Alves Pereira. Os dois foram condenados a 19 anos de prisão cada um. O crime ocorreu em dezembro de 88 AC. A polícia de Minas iniciou as buscas em dezembro, por a polícia, os dois estariam escondidos em uma fazenda de difícil acesso.

6. FSP950111-026

   A juíza Marilene Soares Reis Franco, Vara Federal, acatou denúncia e de a Fazenda, Marco Aurélio é filho de o governador Marcello Alencar PSDB. Os dois foram denunciados por o Ministério Público, acusados de irregularidades secretário municipal de Fazenda. A denúncia foi feita em 6 de dezembro. O governador é uma de as testemunhas arroladas por Miranda Da Sucursal do Rio.

7. FSP950111-034 Pressões políticas fizeram o governo adiar a posse de os presidentes de o Banco do Brasil, e de a Caixa Econômica Federal, A posse de Ximenes estava marcada para hoje, e a de Cutolo, Aliados de o governo vêm reivindicando cargos. A equipe econômica quer que todas as diretorias sejam ocupadas por técnicos. O ministro da Fazenda, decidiu só empossar Ximenes e Cutolo. A equipe econômica decidiu apressar a escolha de as diretorias. Ximenes e Cutolo fizeram as indicações. Os nomes podem ser aprovados até a próxima semana. Da Sucursal de Brasília.

8. FSP950111-036 O novo presidente de a Câmara dos Deputados de os EUA, demitiu em a noite de segunda-feira a historiadora de a Câmara, recém-indicada por ele. O motivo seria seu apoio a um bloqueio de fundos federais para o estudo de o Holocausto em as escolas.

9. FSP950117-048 O custo médio de os 68 produtos que fazem parte de a cesta básica de o paulistano ficou em R$ 99,02, de 0,12 a cesta acumula queda de 3,60. A pesquisa é realizada por o Procon. O grupo alimentação ficou estável terem sido de produtos que compõem o grupo, que acumula redução de 4,46 Os grupos de higiene pessoal e de produtos de limpeza apresentaram aumentos de 1,26 e 0,22, respectivamente. Os produtos foram salsicha avulsa 6,54, biscoito de maisena Triunfo 4,55 e frango resfriado inteiro 1,59. As maiores quedas foram de a carne deprimeira sem osso $ acém 2,51, leite em pó integral Itambé 1,83 e carne de primeira $ coxão mole 0,90. 26 produtos subiram, 15 baixaram de preço e 27 permaneceram estáveis O custo mínimo de a cesta foi de R$ 67,99.

10. FSP950117-074 Maria da Glória Chagas Pereira, 32, foi presa em flagrante de a o tentar matar a facadas sua mãe, 60. A tentativa de homicídio ocorreu em uma casa em o bairro de o Realengo zona norte de o Rio. Seu estado de saúde era regular DP Realengo, Maria da Glória contou que sofreu uma crise nervosa e atacou sua mãe. Ela sofre de problemas mentais e teria se descontrolado Ronaldo Soares.

11. FSP940101-066 A Fundação Civil Casa de Misericórdia, vai assumir a área de a Saúde de o município afirmou o prefeito. A crise de o setor em Franca agravou-se. Não aceitei ainda a decisão de o Botto. Disse Balieiro. A greve de os médicos municipais, iniciou a crise de a Saúde em Franca. O problema agravou-se Da Folha Nordeste o prefeito Ary Balieiro.

12. FSP940101-074 O shopping Center Norte vai sortear uma viagem a o Caribe. Para concorrer, é preciso trocar notas fiscais recebidas durante as compras por cupons. Cada Cr$ 5.000,00 em notas vale um cupom. O prêmio é um cruzeiro por o Caribe. O slogan de a promoção é Que tal catar coquinho no Caribe? Essa é a última viagem sorteada por o shopping. O Center Norte dá como prêmio uma viagem internacional a seus frequentadores. A promoção vale para as compras feitas a partir de segunda-feira. Os cupons serão trocados em o posto de o shopping Da Reportagem Local.

13. FSP940101-079

    A pesquisa Datafolha é um levantamento por amostragem estratificada, realizado entre os dias 16 e 17 de dezembro, foram entrevistadas 1.076 pessoas 646 432 432 432 432 432 432 428. A direção de o Datafolha é exercida por os sociólogos Antonio Manuel Teixeira Mendes e Gustavo Venturi. Da Redação.

14. FSP940101-085 A maior novidade em o setor de os transportes em o município deve ser a introdução de o sistema de catracas eletrônicas, O sindicato de os condutores é contra a medida e ameaçou fazer greves em protesto. Com a catraca eletrônica, a compra de bilhetes poderia ser antecipada O programa de corredores, outra promessa teve as primeiras licitações lançadas em o final de o ano passado. Vai permitir a integração mais rápida de os bairros os corredores devem começar a ser implantados em o final de o ano. Da Reportagem Local.

15. FSP940101-092 O inquérito policial para investigar as denúncias de corrupção em o futebol de o Rio foi instaurado PSDB. Disse o autor de o projeto de criação de a chamada CPI do apito. As investigações de supostas irregularidades em o futebol fluminense se ampliam. Também o Ministério Público do Estado está investigando as denúncias envolvendo o diretor de a Comissão de Arbitros, e o presidente de a Federação, o esquema favoreceria clubes ligados a os dois. A criação de a CPI do Apito foi adiada Dependendo de o calendário, a decisão ficará para fevereiro. Da Sucursal do Rio

16. FSP940101-095 A Copa Hopman é interessante. É um torneio festivo, disputado com muita descontração. Em a edição de o ano passado, a Alemanha deu um passeio, a principal atração passa a ser uma inovação tecnológica, É um grande passo adiante. O novo sistema tem dois desafios. Um de eles, exclusivamente técnico, é demonstrar eficiência. O outro desafio é mais complicado: É pena que John McEnroe desistiu de participar de o torneio. Seria o piloto-de-testes perfeito para o novo sistema. Big

Mac foi o primeiro a testar o sensor de solo em 1987, Revoltado, ficou xingando as linhas de a área de saque Palhaçadas a cena protagonizada por McEnroe é simbólica Quem acompanha os bastidores de o circuito sabe que todos os jogadores odeiam o sensor. É lógico 75 de os entrevistados votou contra o sensor. Todo mundo gosta de vaiar em uma bola duvidosa. só resta torcer para muitas falhas Thales De Menezes. Da Reportagem Local.

17. FSP940101-102 A fisiologia é o estudo de as funções orgânicas de um ser vivo, o homem. O fisiologista pode ser diplomado em ciências biomédicas ou medicina. Sua formação é diferente de o preparador físico, tem diploma de educação física. O trabalho de um fisiologista tem duas etapas. A primeira é avaliar o estágio atlético atleta ou não. são feitos os famosos testes de bicicleta e esteira grandezas como limiar anaeróbico velocidade máxima do que o corpo é capaz de absorver, potência muscular o fisiologista define as linhas de um programa de condicionamento físico. o programa é passado a o preparador físico, A relação entre os dois profissionais é de coordenação e não de subordinação. Da Reportagem Local Md.

18. FSP940101-107 A brasileira Carmem de Oliveira, 28, disse após a prova feminina ter ficado feliz com o 2.° lugar. Queria melhorar um minuto. Melhorei quase quatro, disse. Fez a prova levou 50min31. Em a outra vez, ela perdeu. Disse seu técnico, A edição saiu com a data de o alto errada. A data correta é 1 de janeiro de 1994. Erramos. Da Reportagem Local. Md

19. FSP940101-124 O fotógrafo inglês David Sims, um de os nomes de a nova geração de a fotografia de moda voltada para a realidade, chegou em Londres cerca de Cr$ 500,00 em o bolso, Sims pensou em ser fotógrafo. Queria ser desenhista de quadrinhos. Sims, explica eu trabalho referindo-se a a HQ: minhas fotos são iguais a os quadrinhos, com as proporções erradas. Acho legal David Sims 26 anos. EJ

20. FSP940101-132 O Filho da Pantera Cor de Rosa, estréia. O filme mostra as atrapalhadas aventuras de o filho ilegítimo de o inspetor Closeau, e circuito. A os Cinemas.

Apêndice E

Simbologia do Analisador Palavras

E.1 Siglas e Símbolos do Palavras

symbol	category	examples
S SUBJ	subject sujeito subjekt	Ninguém gosta de chuva. Retomar o controle foi difícil. No seu sonho, a cidade era toda de vidro. Seja quem for. Tem gente morrendo de fome no Brasil. Fugiram do zôo um hipopótamo e um crocodilo.
Od ACC	direct (accusative) object objeto direto (acusativo) direkte (akkusativ) objekt	Liga a luz! Para combater as doenças do inverno, coma vitaminas. Não tem onde morar. Sempre come um monte de folhas.
Oi DAT	dative object objeto indireto pronominal indirekte (dativ) objekt	Deu-lhe um presente. Empreste-me a sua caneta, por favor! Me mostre seu hipopótamo!
Op PIV	prepositional object objeto preposicional præpositionsobjekt	Não me lembro dele. Falamos sobre a sua proposta. Gostava muito de passear ao longo do rio. Não sabe de nada. Pode contar comigo. Chamamos de objeto preposicional complementos indiretos não substituíveis por pronomes adverbiais.
Cs SC	subject complement predicativo do sujeito subjektsprædikat(iv)	Está doente. Está com febre. A moça parece muito cansada. Nadava nua no mar. Andava zangado todo dia.
Co OC	object complement predicativo do objeto objektsprædikat(iv)	O acho muito chato. Tê-lo feito de propósito o faz um delito.
As Ao ADV	argument adverbial complemento adverbial adverbialargument [can be substituted by adverbial pronoun, valency bound, unlike adjuncts]	Durava muito tempo. (As) A jarra caiu no chão. (As) Não mora mais aqui. Mora em São Paulo. (As) Voltamos ao nosso assunto. (As) Mandaram-nos para Londres. (Ao) Costuma custar mais de mil coroas. (As)

Conjunto página 16, Portuguese Syntax, conforme Bick 2000.

Figura E.1: Simbologia/Siglas utilizadas na Gramática Visl – Palavras, 2000.

symbol	category	examples
fA ADVL	adjunct adverbial adjunto adverbial adverbialadjunkt	Sempre comiam cedo. As crianças jogavam no parque. Feito o trabalho temos tempo para mais uma cerveja. Entraram na vila quando amanhecera. O outro dia (fA) fugiu do zôo (As) um hipopótamo.
fApass PASS	passive adjunct agent of passive adjunto do passivo passivadjunkt	Era o herói do dia e foi elogiado pelo chefe do jardim zoológico.
fC PRED	adjunct predicative (subject adjunct) adjunto predicativo prædikativadjunkt	Sempre nada nua. Cansado, se retirou.
fCsta S<	statement predicative (sentence apposition) aposto da oração sætningsprædikativ	Morreu o cachorro da velha, o que muito a entristece.
fCvoc VOK	vocative adjunct constituinte vocativo vokativadjunkt	Me ajuda, Pedro!

Conjunto página 22, Portuguese Syntax, conforme Bick 2000.

Figura E.2: Simbologia/Siglas utilizadas na Gramática Visl – Palavras, 2000.

symbol		category	examples
np	np propp pronp	noun phrase sintagma nominal nominalsyntagme	Era um homem como outro qualquer. (np) A velha avó dormia na rede. (np) Vou fazê-lo eu mesmo. (pronp) O seu nome era Mário Moreno dos Santos. (propp)
ap	adjp advp detp	adpositional phrase sintagma adposicional adpositionssyntagme	As árvores no jardim eram muito velhas. (adjp) Foi um presidente um pouco iconoclasta. (adjp) Nesta saia, parece mais jovem do que as amigas. (adjp) Costuma falar muito devagar. (advp) Ainda hoje vivem de caça e pesca. (advp) Era muito mais velho do que imaginava. (detp)
vp		verb phrase sintagma verbal verbalsyntagme	Ele continua mexendo nas tarefas dos outros. Vem de lhes propor um acordo. Temos que lhe dar mais dinheiro.
pp		prepositional phrase sintagma preposicional præpositionssyntagme	Abriu a janela da sala Gostou do que viu Pedro da Silva Mudamos para São Paulo

Conjunto página 38, Portuguese Syntax, conforme Bick 2000.

Figura E.3: Simbologia/Siglas utilizadas na Gramática Visl – Palavras, 2000.

symbol		category	examples
UTT	STA	utterance	Não faz nada. [statement]
	QUE	enunciado	Já vais embora? [question]
	COM	ytring	Espera! [command]
	EXC		Pobre de mim! [exclamation]
STA		statement enunciado declarativo adsagn	A terra é redonda. Gosta muito de elefantes. Sim vez. Às sete. Obrigado
QUE		question enunciado interrogativo spergsmå	Quera quer uma cerveja? Já foi para o ministério? Quando?
COM		command enunciado imperativo ordre	Pára com isso! Venha pra cá! Fora!
EXC		exclamation enunciado exclamativo adjåb	Deus! Que beleza! Quanta gente!

Conjunto página 97, Portuguese Syntax, conforme Bick 2000.

Figura E.4: Simbologia/Siglas utilizadas na Gramática Visl – Palavras, 2000.

symbol	category	examples
n	noun nome substantiv (nomen)	árvores **n**(F P) um oitavo **n**(<num> M S)
prop	proper noun nome próprio proprium (egenavn)	Estados-Unidos **prop**(M P) Dinamarca **prop**(F S)
adj	adjective adjetivo adjektiv	belas **adj**(F P) terceiros **adj**(<num> M P)
num	numeral numeral numeralia	duas **num**(F P) 17 **num**(<cif> M P)

Conjunto página 107, Portuguese Syntax, conforme 2000.

Figura E.5: Simbologia/Siglas utilizadas na Gramática Visl – Palavras, 2000.

v	v-fin	finite verb verbo finito finit verbum (bojet i tid)	fizessem **v-fin**(IMPF 3P SUBJ)
	v-inf	infinitive infinitivo infinitiv	fazermos **v-inf**(1P)
	v-pcp	participle particípio participium	comprados **v-pcp**(M P) [attributive] tem comprado **v-pcp** [verbal]
	v-ger	gerund gerúndio gerundium	correndo **v-ger**
art		article artigo artikel	os membros **art**(<artd> M P) [definite] uma criança **art**(<arti> F S) [indefinite]
pron	pron-pers	personal pronoun pronome pessoal personligt pronomen	mim **pron-pers**(1S PIV) tu **pron-pers**(2S NOM)
	pron-det	determiner pronoun pronome determinativo determinativt pronomen (adjektivisk pronomen)	estas **pron-det**(<dem> F P) [demonstrative] muita **pron-det**(<quant> F S) [indefinite] cujos **pron-det**(<rel> M P) [relative] quantos **pron-det**(<interr> M P) [interrogative] minhas **pron-det**(<poss 1P> F P) [possessive]
	pron-indp	independent pronoun pronome independente independent pronomen (substantivisk pronomen)	isto **pron-indp**(<dem> M S) [demonstrative] algo, nada **pron-indp**(<quant> M S) [indefinite] os-quais **pron-indp**(<rel> M P) [relative] quem **pron-indp**(<interr> M S) [interrogative]
adv		adverb advérbio adverbium	facilmente, devagar **adv** [modals] aqui, lá **adv** [pronominals] muito, imensamente **adv** [intensifiers] onde, quando, como **adv** [relatives or interrogatives] não, até, já **adv** [operators]
prp		preposition preposição preposition	contra **prp** em-vez-de **prp** <c>
in		interjection interjeição interjektion	oi! **in**
conj	conj-s	subordinating conjunction conjunção subordinativa underordnende konjunktion	que **conj-s** embora **conj-s**
	conj-c	coordinating conjunction conjunção coordenativa sideordnende konjunktion	e **conj-c** ou **conj-c**
pu		punctuation pontuação tegnsætningstegn	, **pu** [komma]

Conjunto página 108, Portuguese Syntax, conforme Bick 2000.

Figura E.6: Simbologia/Siglas utilizadas na Gramática Visl – Palavras, 2000.

Apêndice F

Printscreen do sistema AuTema-Dis

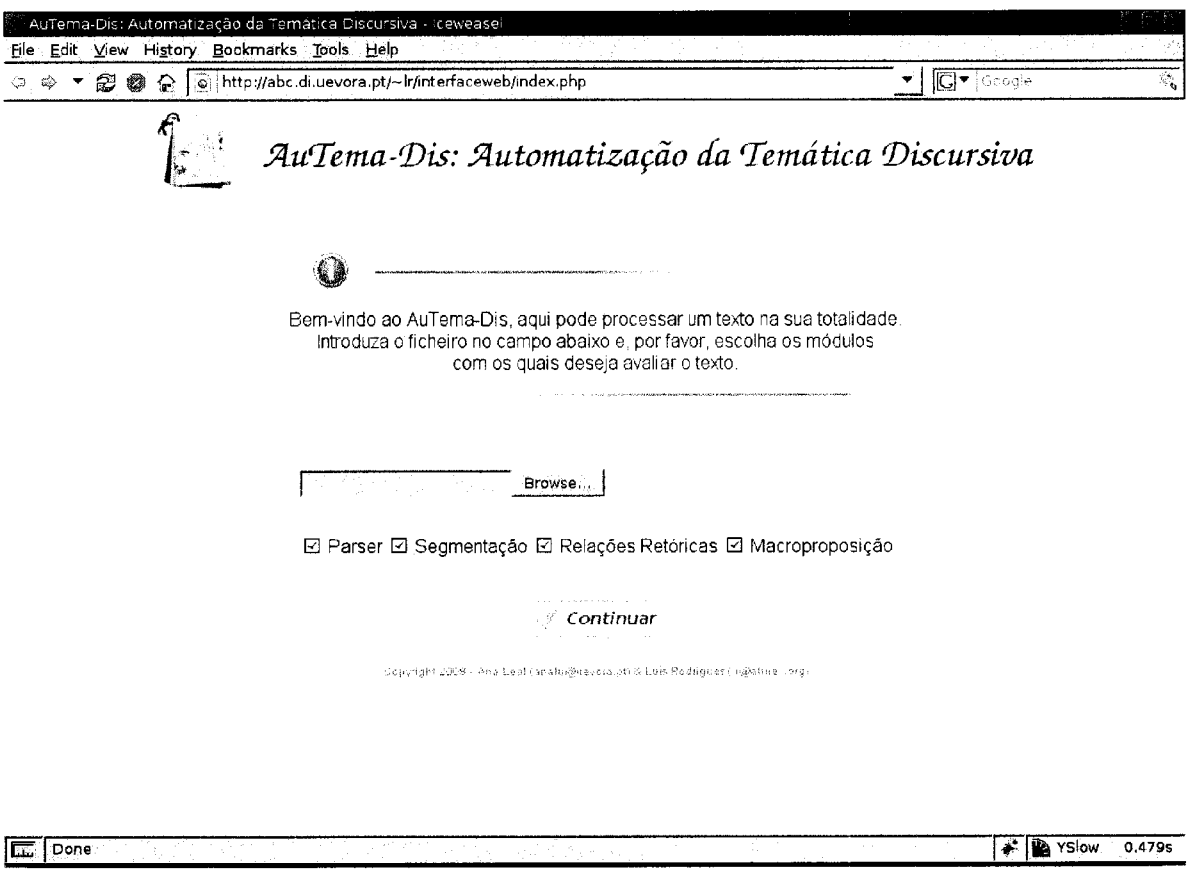

Figura F.1: A figura representa a interface inicial do sistema AuTema-Dis, em que o usuário introduz o texto a ser processado e escolhe as etapas a serem apresentadas.

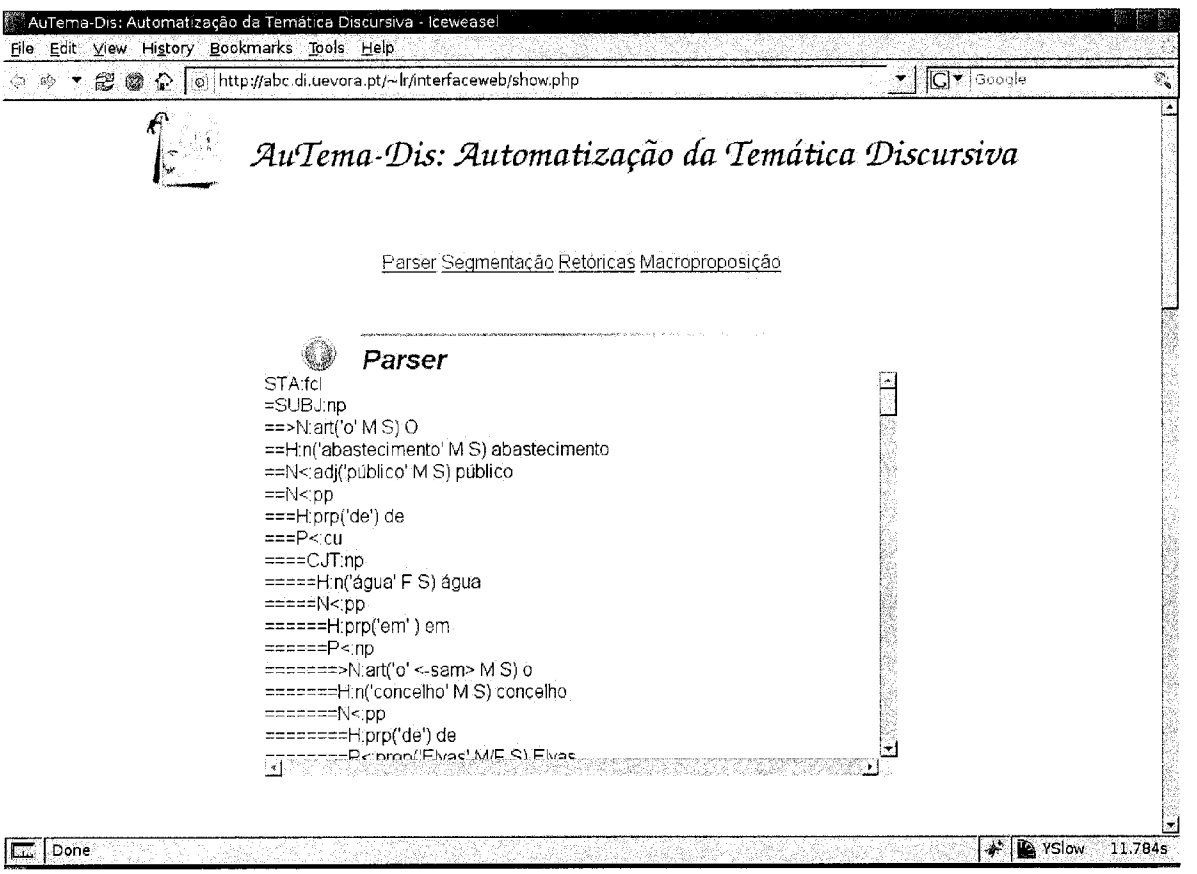

Figura F.2: A figura representa o resultado do processamento do texto selecionado analisado automaticamente pelo analisador *Palavras*.

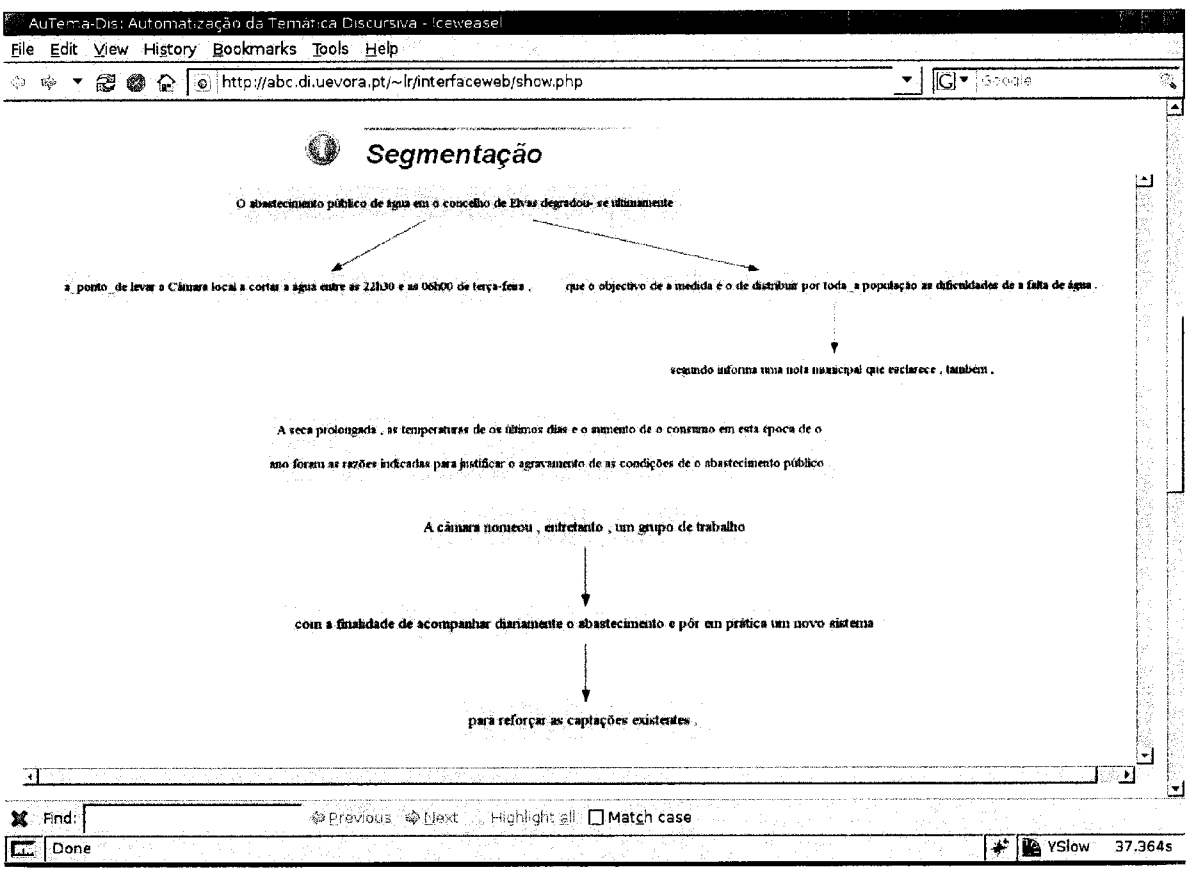

Figura F.3: A figura representa a 2ª etapa realizada pelo AuTema-Dis, a qual segmenta o texto em unidades – segmentos e subsegmentos, organizando-os em árvores de dependência de segmentos – DTS's.

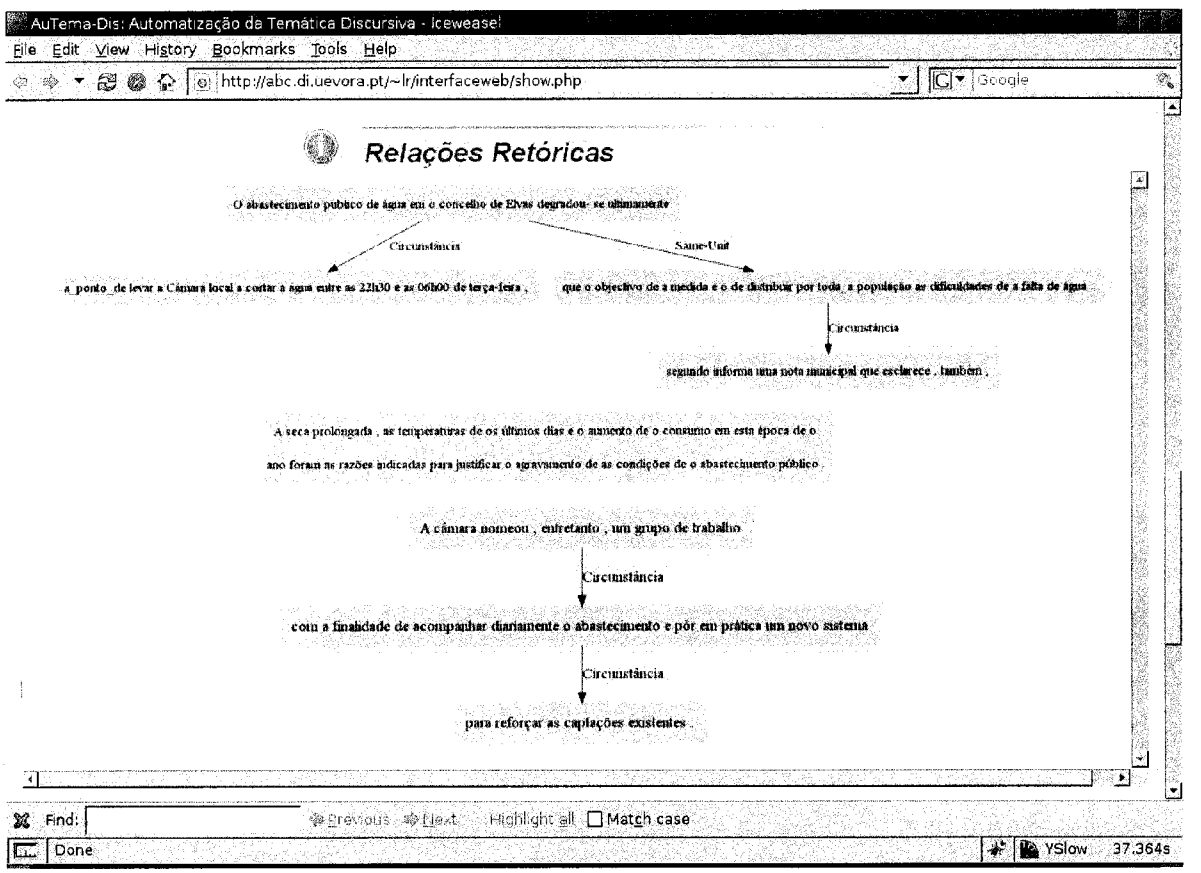

Figura F.4: A figura representa a 3ª etapa realizada pelo AuTema-Dis, em que são atribuídas automaticamente relações retóricas entre os constituintes organizados nas DTS's .

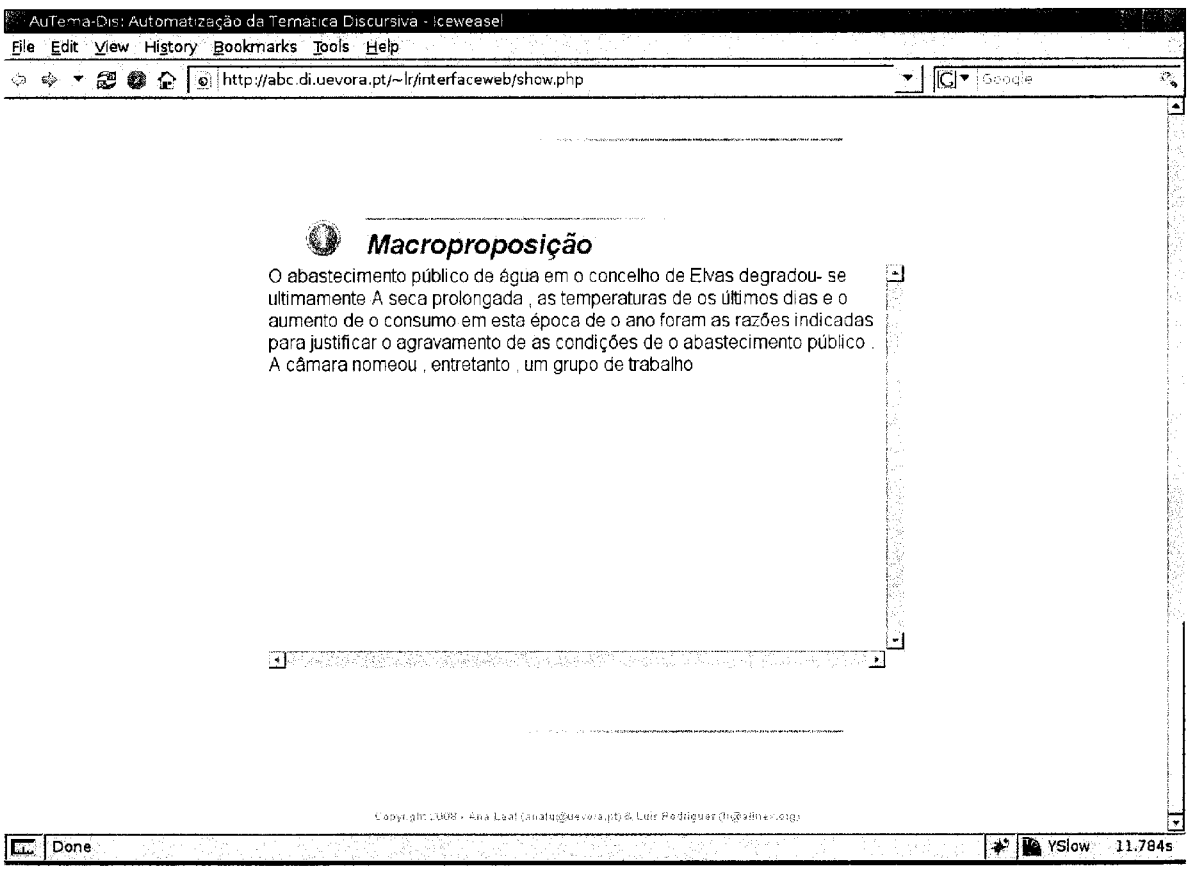

Figura F.5: A figura representa a 4ª etapa realizada pelo AuTema-DIs, na qual o sistema apresenta automaticamente a macroestrutura/macroproposição do texto analisado.

Referências Bibliográficas

[1] E. Bick. *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Aarhus University Press, 2000.

[2] L. Carlson and D. Marcu. Discourse tagging reference manual. Technical Report ISITR545, ISI Technical Report, 2001.

[3] S. Corston-Olivier. *Computing Representations of the Structure of Writen Discourse.* PhD thesis, 1998.

[4] M. da Graça Costa Val. *Redação E Textualidade.* Livraria e Martins Fontes Editora Ltda, 1999.

[5] T. A. V. Dijk. *Some Aspects of Text Grammars.* The Hague:Mouton, 1972.

[6] T. A. V. Dijk. Etudes du discourse et enseignemet. Linguistique et Semiologie, 1980.

[7] T. A. V. Dijk. *Macroestructures.* Hillsdale : Lawrence Erlbaum Associates, 1980.

[8] T. A. V. Dijk. Discourse studies and education. (2):1–26, 1982.

[9] T. A. V. Dijk. *La ciencia del texto.* Barcelona: Paidós, 1989.

[10] T. A. V. Dijk. *Cognição, discurso e interação.* São Paulo: Contexto, 1992.

[11] T. A. V. Dijk. *Texto y contexto. Semántica y pragmática del discurso.* Cátedra : Madrid, 1993.

[12] T. A. V. Dijk and W. Kintsch. *Strategies of discourse comprehension.* New York: Academic, 1983.

[13] L. Favero and I. Koch. *Linguística Textual : Introdução.* Cortes: São Paulo, 1994.

[14] B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. 12(3):175–204, 1986.

[15] F. Jean. Case studies in organizational research. In *Qualitative methods in organizational research: a practical guide*, pages 208–229. London – Sage, 1994.

[16] T. Joachims. *Learning to Classify Text Using Support Vector Machines – Methods, Theory and Algorithms*, volume 668 of *The Springer International Series in Engineering and Computer Science.* Kluwer Academic Publishers/Springer, 2002.

[17] M. Jordan. *An Integrated Three-Pronged Analysis of a Fund-Raising Letter. In W. C.Mann and S. A. Thompson (eds), Discourse Description: Diverse Linguistic Analyses of a Fund–Raising text.* Number 16. John Benjamins Publishing Co, 1992.

[18] W. Kintsch and T. A. V. Dijk. Toward a model of text comprehension and production. (85):363–394, 1978.

[19] A. Knott, M. O'Donnell, J. Oberlander, and C. Mellish. Defeasible rules in content selection and text structuring. In *Proceedings of the 6th European Workshop on Natural Language Generation*, 1997.

[20] I. Koche and L. Travaglia. *A Coerência Textual.* Editora Contexto, 2003.

[21] A. L. Leal and P. Quaresma. Desenvolvimento e integração de recursos para pesquisa de informação – um processo interdisciplinar e interinstitucional. In C. Sarmento, editor, *XVI Encontro da Associação das Universidades de Língua Portuguesa (2006)*, pages 99–108, Macau, China, Fevereiro 2008. Associação das Universidades de Língua Portuguesa.

[22] A. L. Leal, P. Quaresma, and R. Chishman. From syntactical analysis to textual segmentation. In Vieira et al. [43], pages 252–255.

[23] W. Mann and S. Thompson. Rhetorical structure theory: toward a functional theory of text organization. Technical Report Technical Report ISIRS87190, University of Southern California, 1987.

[24] W. Mann and S. Thompson. Rhetorical structure theory: toward a functional theory of text organization. 3(8):243–281, 1988.

[25] D. Marcu. *The Rhetorical Parsing, Summarization and Generation of Natural Language Text Organization.* PhD thesis, 1997.

[26] D. Marcu. Extending a formal and computational model of rhetorical structure theory with intentional structures à la grosz and sidner. In *The 18th International Conference on Computational Linguistics (COLING2000)*, 2000.

[27] D. Marcu. *The Theory and Practice of Discourse Parsing and Summarization.* The MIT Press, Cambridge, MA, 2000.

[28] D. Marcu and A. Echihabi. An unsupervised approach to recognizing discourse relations. In *The Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL02)*, 2002.

[29] D. Marcu, C. Lynn, and W. Maki. The automatic translation of discourse structure. In *1st Meeting of the North Americam Chapter of the Association for Computational Linguistics (NAACL00)*, volume I, pages 09–17, 2000.

[30] L. A. Marcuschi. *Gêneros textuais: definição e funcionalidade. In DIONÍSIO, A. et al. Gêneros textuais e ensino.* Lucerna, Rio de Janeiro, 2002.

[31] S. C. Marquesi. *A organização do texto descritivo em língua portuguesa.* Rio de Janeiro, 2 edition, 2004.

[32] C. Matthiessen, M. O'Donnell, and L. Zeng. Discourse analysis and the need for functionally complex grammars in parsing. In *Proceedings of the Second Japan – Australia Joint Symposium on Natural Language Processing*, 1991.

[33] M. O'Donnell. *Sentence Analysis and Geration – a Systemic Perspective.* PhD thesis, 1994.

[34] M. O'Donnell. Rst – tool: An RST analysis tool. In *Proceedings of the 6th European Workshop on Natural Language Generation*, 1997.

[35] M. O'Donnell. Variable – length on–line document generation. In *Proceedings of the 6th European Workshop on Natural Language Generation*, 1997.

[36] M. O'Donnell. Intermixing multiple discourse strategies for automatic tex composition. (40), 2000.

[37] M. O'Donnell. RSTTool 2.4 – a markup tool for rhetorical structure theory. In *Proceedings of the International Natural Language Generation Conference (INLG2000)*, 2000.

[38] T. Pardo. *Métodos para Análise Discursiva Automática.* PhD thesis, 2005.

[39] T. Pardo and M. da Graça Nunes. Análise de discurso – teorias discursivas e aplicações em processamento de línguas naturais. Technical Report 196, Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo, 2003.

[40] R. Soricut and D. Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL)*, 2003.

[41] M. Taboada and W. Mann. Applications of rhetorical structure theory. Technical Report Technical Report, University of Southern California, 2005.

[42] M. Taboada and W. Mann. *Rhetorical Structure Theory: looking back and moving ahead.* Number 8. Discourse Studies – Sage Publications, 2006.

[43] R. Vieira, P. Quaresma, M. G. V. Nunes, N. Mamede, C. Oliveira, and M. C. Dias, editors. *Computational Processing of the Portuguese Language, Propor 2006, Itatiaia, Brasil, May 13-17, 2006, Proceedings*, volume 3960 of *Lecture Notes in Computer Science.* Springer, 2006.

[44] L. S. Vygotskij. *Thought and language.* Cambridge: MIT Press, 1962.