UNIVERSIDADE FEDERAL DE SÃO CARLOS

DEPARTAMENTO DE LETRAS

PROGRAMA DE PÓS-GRADUAÇÃO EM LÍNGÜÍSTICA

Estudo e validação de teorias do domínio lingüístico com vistas à melhoria do tratamento de cadeias de co-referência em Sumarização Automática

Thiago Ianez Carbonel

Orientadora: *Dra. Lucia Helena Machado Rino*

São Carlos

Agosto de 2007

UNIVERSIDADE FEDERAL DE SÃO CARLOS

DEPARTAMENTO DE LETRAS

PROGRAMA DE PÓS-GRADUAÇÃO EM LÍNGÜÍSTICA

Estudo e validação de teorias do domínio lingüístico com vistas à melhoria do tratamento de cadeias de co-referência em Sumarização Automática

Thiago Ianez Carbonel

Dissertação apresentada ao Departamento de Letras da Universidade Federal de São Carlos – DL/UFSCAR, como parte dos requisitos para obtenção do título de Mestre em Lingüística.

São Carlos

Agosto de 2007

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

C264ev

Carbonel, Thiago Ianez.
Estudo e validação de teorias do domínio lingüístico com vistas à melhoria do tratamento de cadeias de co-referência em Sumarização Automática / Thiago Ianez Carbonel. -- São Carlos : UFSCar, 2007.
189 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2007.

1. Lingüística – processamento de dados. 2. Textualidade. 3. Sumarização Automática. 4. Anáfora (Lingüística). I. Título.

CDD: 410.285 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Lingüística

Rodovia Washington Luis, Km 235 - Caixa Postal 676
CEP: 13565-905 – São Carlos – São Paulo – Brasil
Telefone (16) 3351-8360 - Fax: (16) 3351-8353
ppgl@power.ufscar.br www.ppgl.ufscar.br

PPGL
Programa de Pós-Graduação em Lingüística

ATA DO EXAME DE DEFESA DA DISSERTAÇÃO DE MESTRADO DE TIAGO IANEZ CARBONEL

Área de Concentração: Estudos Lingüísticos
Linha de Pesquisa: Linguagem Humana e Tecnologia

Aos **21** dias do mês de **agosto** do ano de **dois mil e sete**, às **16 horas**, na **Sala de de Defesa do PPGE** da Universidade Federal de São Carlos, reuniu-se a Banca Examinadora nas formas e termos do artigo 23º do Regimento Interno do Programa de Pós-Graduação em Lingüística, com a seguinte composição: **Profª. Dra. Lucia Helena Machado Rino (UFSCar/São Carlos – Orientadora/Presidente)**, **Profª. Dra. Renata Vieira (INISINOS/RS – Membro Titular)** e **Prof. Dr. Thiago Alexandre Salgueiro Pardo (ICMC-USP/São Carlos - Membro Titular),** para o exame de defesa da **Dissertação** de **Mestrado** de **Thiago Ianez Carbonel**, intitulada: **"Estudo e validação de teorias do domínio lingüístico com vistas à melhoria do tratamento de cadeias de co-referências em Sumarização Automática"**. A sessão pública foi instalada pela Presidente da Banca Examinadora, a qual, após explanação da candidata, passou a palavra aos demais membros da Banca. Terminada a argüição, a Banca Examinadora reuniu-se em sessão secreta, tendo atribuído à candidata o(s) conceito(s): A, por unanimidade.

De acordo com o artigo 23º, o candidato foi aprovado. Nada mais havendo a tratar, foi encerrada a sessão e para constar, eu, Nancira Marinho Ribeiro, Assistente Administrativo do PPGL, lavrei a presente ata assinada pelos membros da Banca Examinadora.

São Carlos, 21 de agosto de 2007.

Profª. Dra. Lucia Helena Machado Rino
Profª. Dra. Renata Vieira
Prof. Dr. Thiago Alexandre Salgueiro Pardo

É difícil defender só com palavras a vida.

João Cabral de Melo Neto

AGRADECIMENTOS

Primeiramente, aos meus pais pelo amor, apoio e fé – sem vocês eu teria parado bem antes daqui.

A minha avó, Elda (*in memoriam*), pelo amor incondicional e pela crença em mim.

A minha orientadora, Lucia, por tudo, desde a atenção total que sempre devotou ao meu trabalho até os "puxões de orelha" que, como boa "mãe postiça", me deu ao longo do desenvolvimento da pesquisa e da escrita, colocando-me nos trilhos corretos.

Aos meus colegas do NILC (tanto da matriz USP quanto da "sucursal" UFSCAR), pelo apoio nas horas de dúvida, pelos conselhos e orientações, pela amizade, pelo tempo que puderam me dedicar. Agradeço especialmente àqueles que, como Elô, Daniel e Thiago, se envolveram diretamente na minha pesquisa, fornecendo-me material, apoio e tudo que precisei.

Aos grandes amigos Thiago, Ariani e Anselmo que, particularmente, estiveram próximos o tempo todo, tanto nas discussões de trabalho, quanto na *happy hour*.

Aos colegas da UNISINOS – Renata, Sandrinha, Juliana – que, mesmo de longe, contribuíram com o meu trabalho.

A minha grande e especialíssima amiga-musa Ira, pois sem ela meus caminhos nessa vida teriam sido outros.

E, por fim, mas com uma imensa importância, ao Jorge, companheiro na vida e em todo o resto, pela ajuda e atenção, pela devoção e empenho – esse trabalho é mais nosso do que apenas meu.

Às agências de fomento à pesquisa CNPq e CAPES.

RESUMO

O trabalho apresentado nesta dissertação tem como foco o estudo e validação de teorias lingüísticas com vistas à melhoria dos sistemas de Sumarização Automática, ramo da Lingüística Computacional que, com o advento da internet, tem recebido grande atenção nos últimos tempos, pois a quantidade de informação *on-line* é enorme e os leitores têm cada vez menos tempo para apreender o máximo desta informação.
Nesta dissertação, avaliamos o protótipo de AS baseado na Teoria das Veias proposto por Seno (2005) e apresentamos uma reimplementação com características distintas, baseada em análise de córpus – um córpus anotado com informação retórica (RST) e referencial. Como inovação metodológica, formulamos a *Precisão Não-Trivial*, um estimador mais realista para o poder preditivo da C1.

ABSTRACT

The work presented in the dissertation focuses on the study and validation of linguistic theories so as to improve reference cohesion in Automatic Summarization systems, which with the advent of the Internet have received increasing attention due to the urge to manage the huge amounts of on-line textual information that become available each day.
In this dissertation we evaluate Seno (2005)'s Veins Theory-based proposal and prototype, and present a reimplementation with distinct features based on the analysis of a corpus annotated with rhetoric (RST) and referential information. In addition, we report on the first validation effort for Portuguese for Veins Theory's Conjecture 1 (C1), which constrains anaphora resolution given the rhetoric structure of texts and whose applicability to Automatic Summarization interests us. As a methodological novelty, we put forth the Non-Trivial Precision*, a more realistic estimator of C1's predictive power.*

ÍNDICE GERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

1. Introdução

Devido à grande quantidade de informação textual disponível atualmente, sobretudo na Internet, sistemas capazes de condensar esses textos são altamente necessários, pois diminuem o descompasso entre sua produção e sua absorção. Tais sistemas – sistemas de Sumarização Automática (SA) – partem do pressuposto de que é possível gerar automaticamente um sumário[1] a partir de um dado texto. Trabalhos relevantes têm sido apresentados nessa área (Sparck Jones, 1993; Mani, 2001; Mani & Maybury, 1999; Pardo & Rino, 2001; Pardo et al., 2003; Seno, 2005; Seno & Rino, 2005).

Entretanto, ainda não dispomos de uma modelagem satisfatória para garantir a textualidade[2] dos sumários automáticos. Os sistemas de SA apresentam dificuldades de processamento ao lidarem com determinados fenômenos lingüísticos, tais como as cadeias de co-referência e a ambigüidade lexical.

Quando se fala em modelagem satisfatória, é preciso ter em foco o objetivo da modelagem, o que, no caso, é fornecer subsídios para o processamento algorítmico – forma de "pensar" da máquina. O homem, usuário e manipulador natural da língua, processa estes fatos da linguagem cognitivamente, utilizando sua inteligência, ou seja, utiliza um conjunto de processos mentais que são de difícil reprodução computacional por não serem passíveis de formulação algorítmica (Pelizzoni, 2005). Parte das dificuldades enfrentadas na área da Lingüística Computacional guarda estreita relação com a questão da formulação algorítmica. Sistemas automáticos de processamento de língua natural demandam descrições de fenômenos lingüísticos que possibilitem suas representações computáveis dos mesmos. As línguas naturais, porém, apresentam fenômenos que se processam através de elaboradas construções mentais que não se traduzem em informação matematicamente computável – e um exemplo bastante claro disso é a própria idéia de metáfora. O cérebro

[1] Utilizamos o termo sumário como sinônimo de resumo, mesmo considerando as sutis diferenças de significado entre os termos. Tal escolha já foi cristalizada na área da Lingüística Computacional – Sparck-Jones (1993).
[2] A idéia de textualidade pode ser compreendida como o conjunto de características que fazem com que um texto seja um texto, e não apenas uma seqüência de frases (Costa Val, 1991).

humano estabelece intrincadas relações semânticas para a criação da linguagem figurada e à Lingüística, até então, interessava apenas a descrição do fenômeno, ainda que para tanto fosse necessária a abstração do processo mental, descrito apenas como relações cognitivas. Para a máquina, no entanto, essa descrição não tem serventia, uma vez que é preciso o dado concreto, manipulável e passível de processamento.

Temos abaixo um exemplo[3] de texto-fonte e sumário produzido, respectivamente:

Texto-fonte: O primeiro-ministro britânico, Tony Blair, admitiu que o protocolo de Kyoto não está funcionando e pediu um novo acordo internacional para combater o aquecimento global.
Em artigo publicado neste domingo pelo jornal The Observer, de Londres, Blair disse que, para que um novo acordo funcione, precisa incluir os Estados Unidos, que são os maiores emissores de gases poluentes do mundo, mas optaram por não ratificar o protocolo de Kyoto.
Os comentários foram feitos antes de uma conferência sobre mudanças climáticas em Londres na terça-feira.

Sumário: O primeiro-ministro britânico, Tony Blair, admitiu que o protocolo de Kyoto não está Em artigo publicado neste domingo pelo jornal The Observer, Os comentários foram feitos antes de uma conferência sobre mudanças climáticas em Londres na terça-feira.

No texto-fonte, o sintagma nominal (SN) os comentários recupera todo o primeiro parágrafo, através de um fenômeno descrito adiante como **encapsulamento** (seção 6)**.** O sumário correspondente não recupera toda a informação do antecedente, uma vez que contém apenas parte do referido parágrafo. Desse modo, o que temos é um sumário deficiente no que se refere, entre outros aspectos, à resolução anafórica e, conseqüentemente, com problemas de textualidade.

Os sistemas de SA disponíveis hoje fazem a identificação da informação mais relevante de um texto, possibilitando a seleção dos segmentos textuais necessários para a síntese de um sumário, como é o caso do GistSumm (Pardo, 2002). Tais sistemas, porém, não garantem que os sumários gerados serão estruturas textuais coesas e coerentes. Entre os problemas que podem ser observados, damos destaque especial à preservação das cadeias de co-

[3] O sumário foi produzido pelo sumarizador automático RheSumaRST (Seno, 2005), e os resultados foram descritos em Carbonel & Rino (2006).

referência e, conseqüentemente, à resolução anafórica. Abundam trabalhos concebidos no universo da língua inglesa (por exemplo Azzam et al., 1999; Mitkov, 2002), mas, para a Língua Portuguesa, há uma grande carência de modelos de resolução anafórica. A despeito dos esforços existentes, não dispomos ainda de um modelo que descreva satisfatoriamente o fenômeno da co-referenciação nessa língua.

O trabalho que se desenvolve neste projeto é um estudo de viés lingüístico acerca dos problemas de textualidade apresentados por sumários gerados automaticamente, bem como uma investigação dos modelos e teorias já existentes na literatura específica – desenvolvidos no âmbito de outras línguas, particularmente o inglês – a fim de avaliar a viabilidade de aplicação em modelos práticos para o português.

No contexto ora descrito, o foco do projeto é uma análise lingüística de resultados gerados automaticamente por sistemas computacionais de sumarização existentes no NILC[4]. Os sistemas que utilizamos (RheSumaRST (Seno, 2005) e o VeinSum - protótipo de SA que apresentamos neste trabalho[5]) remetem a programas que mimetizam o processamento de língua natural feito por humanos – razão pela qual são denominados sistemas de processamento de línguas naturais (PLN)[6]. Assim, fica evidenciado que lidaremos ao longo deste trabalho com a dicotomia entre o "processador humano" e o "processador automático", sendo que constantemente deveremos resgatar o objetivo de automação neste trabalho.

Nesse diapasão, a contribuição deste trabalho é fornecer um estudo aprofundando dos fenômenos textuais que ocorrem nos sumários gerados automaticamente e que são responsáveis por déficits de textualidade, em especial os referentes a quebras de elos co-referenciais. Assim, um objetivo central neste projeto foi o estudo da Teoria de Estruturação Retórica (RST – Mann & Thompson, 1987) – a fim de buscar possíveis relações entre a estrutura retórica do texto e a construção dos elos coesivos – aliado ao estudo e avaliação da Teoria das Veias (Cristea et al., 1998) – que propõe um modelo de

[4] NILC – **N**úcleo **I**nterinstitucional de **L**ingüística **C**omputacional – site: www.nilc.icmc.usp.br
[5] Projeto e Desenvolvimento em parceria com Jorge Marques Pelizzoni.
[6] Ou PALN (Processamento Automático de Línguas Naturais), como prefere Dias-da-Silva (1996).

mapeamento do fenômeno referencial a partir da estruturação retórica baseada na RST. Derivados desse objetivo principal, foram nossos objetivos específicos: i) a construção de córpus de textos analisados retoricamente com vistas a análises da RST e de sua aplicabilidade na Sumarização Automática; ii) anotação automática das veias nesses córpus e conseqüente validação da Teoria das Veias para o português; iii) análises de sumários automaticamente produzidos a fim de identificar e definir as quebras de cadeias de co-referência, possibilitando a construção do conceito de quebra.

O presente projeto de mestrado insere-se no contexto de pesquisa do Projeto PRoCaCoSa (PROcessamento de CAdeias de CO-referência em Sumarização Automática de Textos em Português do Brasil), desenvolvido em parceria com a UNISINOS, e financiado pelo CNPq.

Esta dissertação de mestrado divide-se em duas partes: uma parte teórica, na qual introduzimos os conceitos e teorias essenciais ao trabalho desenvolvido, e uma parte prática, na qual descrevemos os córpus utilizados e sua anotação, bem como os experimentos, os resultados e as discussões que constituem a verdadeira contribuição da pesquisa realizada. Desse modo, na seção 2 tratamos dos conceitos básicos de Sumarização Automática, seguidos, na seção 3, pelos conceitos-chave da Lingüística Textual. Na seção 4, apresentamos modelos lingüístico-computacionais de processamento de textos e, na seção 5, introduzimos os modelos teóricos de representação do conhecimento lingüístico que são a base do estudo desenvolvido neste trabalho. Na parte prática, temos, nas seções 6 e 7, a descrição da anotação do córpus com cadeias de co-referência e com estruturas retóricas (RST). Na seção 8, descrevemos os sistemas automáticos utilizados nos experimentos descritos na seção 9. Os apêndices A e B e C trazem, respectivamente, o conjunto das relações retóricas utilizadas nos experimentos, os doze textos selecionados para a análise neste trabalho e seus sumários automaticamente produzidos pelos sistemas com os quais trabalhamos.

Parte I. Suporte teórico ao problema tratado

2. Sumarização automática

O ato de sumarizar é essencialmente humano e remete a processos cognitivos por meio dos quais o leitor de um texto produz, mentalmente, uma versão simplificada do mesmo, um novo texto de menores dimensões e que contém a informação mais relevante presente no texto original. A máquina, ao produzir um sumário, processa o texto analogamente.

Segundo Rino & Pardo (2003, p. 1), um sumarizador automático tem por objetivo "produzir uma representação condensada do conteúdo mais importante de um texto de entrada, para consumo por usuários humanos. Para isso, ele deve ser capaz de identificar, em um texto ou em uma representação conceitual do mesmo, o que é relevante, estruturando as unidades informativas correspondentes de modo a assegurar que o sumário seja coerente e consistente", ou, nos termos da Lingüística Textual (Koch, 2004), garantir a textualidade do sumário.

Essa caracterização é estritamente voltada à formulação de modelos computacionais, já que se refere a representações condensadas e entradas de sistemas computacionais, e sugere um tratamento diferenciado da tarefa de sumarização de textos quando comparado à acepção lingüística ou, mais precisamente, à tarefa humana correspondente.

Sparck Jones (1993) descreveu a sumarização automática como um processo que se desenvolve, basicamente, em duas etapas: (1) a construção de uma representação do texto-fonte e (2) a geração de uma estrutura condensada a partir desta representação (o sumário é gerado ainda sob a forma desta representação) e sua síntese em língua natural. Nesse sentido Mani & Malbury (1999) apontam uma arquitetura geral dos sistemas de sumarização automática, conforme a Figura 1 (Rino & Pardo, 2003 *apud* Mani & Maybury, 1999).

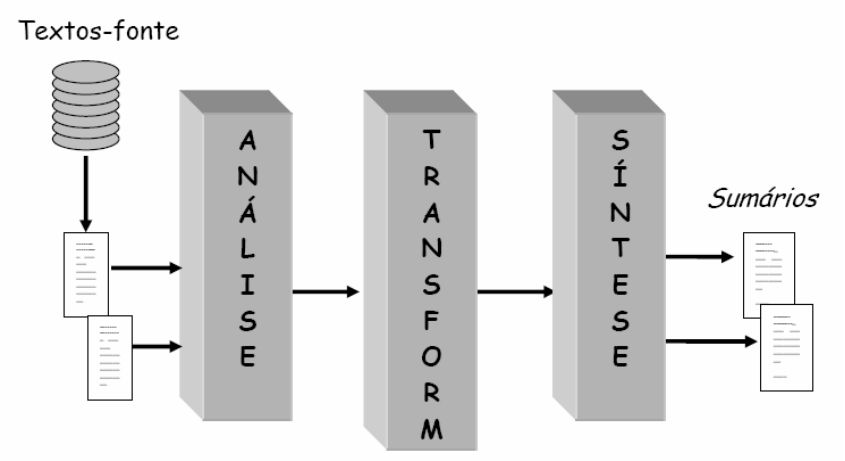


Figura 1. Arquitetura geral de um sistema de SA

Na figura acima, o processo de análise corresponde à interpretação do texto-fonte, gerando, assim, uma representação do conhecimento lingüístico expresso em termos computáveis. A transformação é a etapa em que esse conhecimento é manipulado, gerando o sumário ainda sob a forma de uma representação computável, ou seja, não-textual. E na tarefa de síntese, por fim, o sumário ganha a forma textual.

A partir desta arquitetura, então, podemos determinar que os sistemas de Sumarização Automática correspondem a modelos computacionais capazes de (a) manipular dados textuais; (b) elaborar um processamento automático pré-determinado; (c) produzir resultados esperados ou satisfatórios. Esta é uma visão do processo, mas é preciso atentar às premissas que subjazem ao mesmo.

As principais premissas da sumarização – automática ou não – podem, assim, ser enumeradas como segue (Rino & Pardo, 2003):

- Está disponível um texto, aqui denominado *texto-fonte*, que deve ser condensado.
- A afirmação de que o objeto a ser sumarizado constitui um texto implica, adicionalmente, a existência de:

    a) uma idéia central – o tópico principal do texto – sobre a qual se constrói a trama textual (no ensino fundamental, aprendemos que o texto deve ser desenvolvido a partir de uma idéia – nossa idéia central);

b) um conjunto de unidades de informação que, reconhecidamente, têm relação com a idéia central em desenvolvimento;

c) um objetivo comunicativo central que, implícita ou explicitamente, direciona tanto a seleção das unidades de informação quanto a seleção da forma como a informação será estruturada, para estabelecer a idéia (mensagem) pretendida.

d) um enredo, tecido em função das escolhas antes citadas, visando transmitir a idéia central de forma coerente, a fim de atingir o objetivo comunicativo pretendido.

- Tomando por base essa relação de conceitos, a principal premissa da sumarização de textos pode ser, assim, expressa como a tarefa de *identificar o que é relevante* no texto e, então, *traçar o novo enredo*, a partir do conteúdo disponível, *preservando sua idéia central*, sem transgredir o significado original pretendido.

- A não-transgressão do original constitui a principal restrição da sumarização[7].

Em se tratando do processo cognitivo de sumarização, as premissas acima são ordenadas e realizadas mentalmente. Considerando, porém, a tarefa automática de sumarização, são muitas as dificuldades enfrentadas pela máquina na organização e realização destas tarefas.

Um sumário, em geral, é gerado a partir de um texto que se apresenta em sua estrutura textual pura, ou seja, uma construção de segmentos e parágrafos a partir da união de palavras. Sistemas de SA, porém, podem ter, além do texto puro, outras formas de entrada que correspondem a representações do texto baseadas em determinado conhecimento lingüístico específico, tal como a estrutura retórica (árvores RST). Em verdade, é na forma de representação do texto que a Lingüística Computacional busca soluções para a descrição e processamentos dos fenômenos lingüísticos, daí a importância de, paralelamente ao sistema de processamento em si, trabalhar-se uma representação consistente, livre de ambigüidade (talvez uma das grandes metas na área) e computável do texto.

[7] Essa premissa é válida para a SA nos moldes do que se tem feito atualmente no âmbito de nossos trabalhos com modelos que extraem do texto-fonte fragmentos para a composição do sumário. No entanto, para a "Sumarização Crítica" esta premissa não é verdadeira, pois um sumário crítico pode inserir informação processada a partir da idéia central do texto sumarizado.

Ao longo dos últimos quarenta anos de pesquisa na área, os estudos apontaram para duas abordagens distintas que se desenvolveram paralelamente. Tais abordagens diferem, principalmente, quanto ao nível de conhecimento profundo envolvido, e também na medida em que contemplam, em diferentes graus, a questão da reescrita textual. A essas duas abordagens costuma-se denominar "metodologia profunda" e "metodologia superficial", que veremos a seguir.

As distinções entre as abordagens de sumarização automática ocorrem no modo como essas etapas do sistema são tratadas – se envolvem pouco ou nenhum conhecimento lingüístico, ou se são ricas no mesmo. Nas seções abaixo, descreveremos, detalhadamente, as referidas abordagens.

2.1 Abordagem superficial

Já na década de 50, surgiu a idéia de que seria possível aplicar técnicas estatísticas combinadas com conhecimento lingüístico superficial a fim de gerar sumários de textos. O trabalho pioneiro de Luhn (1958) baseava-se na idéia de que seria possível, através do cálculo da freqüência das palavras e sentenças no texto, atribuir pesos aos componentes textuais e, a partir daí, fazer a seleção das mais relevantes, gerando assim uma forma de sumário. Esse processo, apesar de envolver algum conhecimento lingüístico, o faz em um nível muito superficial, sendo, obviamente, mais amparado por métodos estatísticos.

O trabalho de Luhn, no entanto, era apenas o começo. Nas décadas seguintes, novos trabalhos refinaram a visão sobre o assunto, lançando novas perspectivas de exploração baseadas no uso de ferramentais capazes de auxiliar no processo de seleção de informação mais relevante, tais como o reconhecimento de "palavras e expressões sinalizadoras" ("antes de mais nada", "muito importante", "a mais relevante" etc.), análise da posição da sentença no texto (frases iniciais apresentam maior saliência, por exemplo), restrição do gênero e domínio (para cada gênero e domínio textuais existe um padrão a ser observado no processamento), uso de dicionários eletrônicos, entre outros (Rino & Pardo, 2003).

A sumarização superficial não busca a simulação do processamento humano. Embora mantenha o mesmo objetivo, qual seja o de produzir textos condensados a partir de uma

fonte textual (e, portanto, sujeito às restrições antes expostas), os métodos empregados envolvem a manipulação de segmentos textuais, em geral, no nível morfológico ou lexical, derivando, via de regra, distribuições de freqüência ou classificações das informações, a fim de determinar porções do texto a serem extraídas diretamente que sejam relevantes para a construção do sumário.

Considerando, então, que a abordagem empírica gera extratos, é notado que o sistema não envolve a reescrita textual, ou seja, o sistema não re-elabora as informações selecionadas como relevantes a partir do texto-fonte. O método superficial baseia-se, principalmente, em modelos de distribuições de freqüência a fim de ranquear os segmentos mais significativos (os referidos excertos) para, ao final, uni-los sob a forma de um alinhamento de entidades textuais que, não raro, não possui atributos textuais – coesão e coerência, principalmente (pontos que veremos adiante).

Observemos a arquitetura de um sistema de SA superficial, apresentada na Figura 2, (Rino & Pardo, 2003 *apud* Mani & Maybury, 1999).

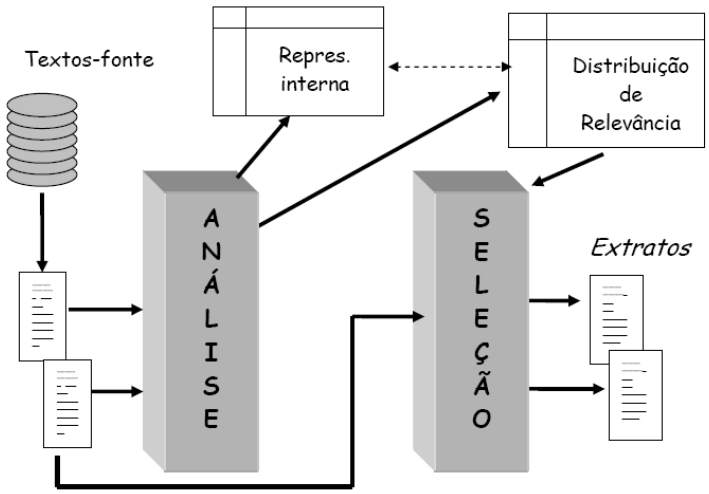


Figura 2. Arquitetura de um sistema de SA superficial

Em geral, para que o sistema faça a seleção através dos métodos extrativos, o texto-fonte é "tokenizado"[8] e etiquetado de acordo com suas características morfológicas e sintáticas. Existem, hoje, vários recursos utilizados para processar um texto-fonte do modo a favorecer a identificação e seleção dos segmentos relevantes para compor o extrato. Esses recursos são variáveis e são adotados diferentemente para cada sistema. Podem ser usados, por exemplo, recursos de remoção de *stopwords* ou de *stemização*. No primeiro caso, deve ser fornecida uma *stoplist*, isto é, um conjunto de *stopwords* (palavras que são consideradas irrelevantes). Tratam-se, usualmente, de palavras muito freqüentes (ou seja, provavelmente ortogonais aos diversos gêneros textuais ou classes lingüísticas de interesse) e de baixo valor semântico, como artigos, preposições, pronomes, alguns advérbios etc. No segundo caso, busca-se a forma canônica (ou radical, ou *stem*) de cada palavra que não esteja na *stoplist*. Em ambos os casos, busca-se melhorar o tratamento superficial do texto. A remoção de *stopwords*, nesse caso, visa impedir que as altas freqüências das mesmas no texto atrapalhem a pontuação dos termos relevantes mais freqüentes. A stemização, por sua vez, visa permitir que o sistema correlacione itens lexicais que não seriam relacionados, caso contrário.

Na Sumarização Automática superficial, o processo de síntese resume-se à justaposição dos segmentados selecionados de acordo com sua relevância, e na medida da taxa de compressão determinada.

2.2 Abordagem profunda

A sumarização automática profunda busca a mimetização – ou uma aproximação disso - , ou melhor, a modelagem do processo humano de sumarização. Um humano, diante das tarefas de ler e compreender um texto (análise), condensar seu conteúdo (processo de sumarização em si) e gerar um sumário (síntese), desempenha-as através de processos

[8] A **tokenização** é um estágio do processamento de um texto de entrada que o converte de uma seqüência de caracteres em uma seqüência de unidades mais significativas para o estágio seguinte, ditas **tokens**. Um tokenizador resolve problemas de baixo nível, tais como se um dado ponto é marca de fim de sentença ou pertence a uma abreviatura, ou ainda se uma dada vírgula é sinal de pontuação ou parte de um número. Por exemplo, a seqüência de caracteres "Ontem, o Dr. Pedro pagou R$ 50,99." poderia ser tokenizada como: <'ontem' – ',' – 'Dr.' – 'Pedro' – 'pagou' – 'R$' – '50,99' – '.'>. Cumpre notar que a definição de token – e, portanto, as especificidades da função do tokenizador – pode variar de sistema para sistema.

cognitivos, reformulando o texto-fonte a fim de elaborar sua versão reduzida. Esse processo de reescrita, ou a mera composição a partir do texto-fonte, envolve um alto grau de conhecimento lingüístico – reformulação de estruturas sintáticas, escolhas lexicais, uso de sinonímia etc.

Nessa abordagem, a máquina, ao fazer o mesmo processo, precisa de uma arquitetura refinada de processos automáticos que seja capaz de "imitar" tais tarefas de maneira algorítmica, envolvendo, por óbvio, um alto nível de conhecimento profundo. Atualmente, existem trabalhos significativos sendo desenvolvidos e um sistema pode ser mencionado como exemplo dessa abordagem: o RheSumaRST (Seno, 2005; Seno & Rino, 2005).

A Figura 3 (Rino & Pardo, 2003 *apud* Mani & Maybury, 1999) descreve, genericamente, a arquitetura de um sistema de SA profunda. Observemos que o processo pretende simular o comportamento humano para a mesma tarefa, ou seja, o computador parte de uma representação conceitual do texto para, após o processamento (redução), ter como saída uma outra representação conceitual. O ser humano, do mesmo modo, ao ler um dado texto *T*, assimila-o sob a forma de um conjunto *X* de informações que, após a seleção (baseada em relevância, por exemplo), reduz-se a um subconjunto *Y* de informações relevantes que deverão ser reagrupadas sob a forma de um texto condensado *S* (sumário resultante).

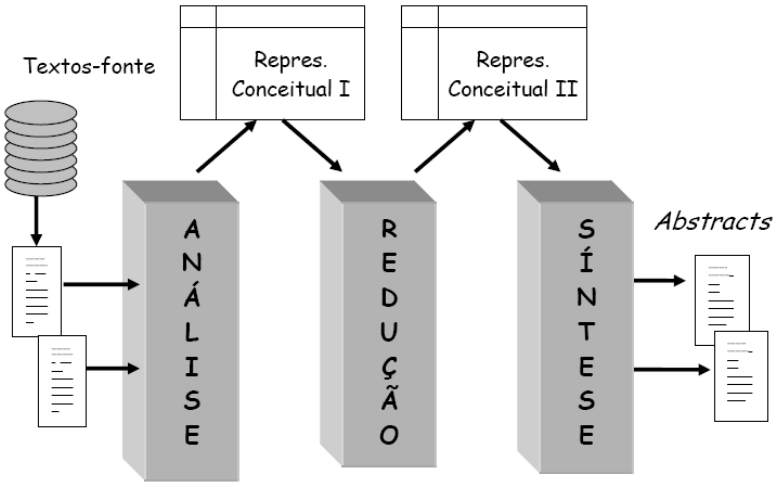


Figura 3. Arquitetura geral de um sistema de SA profunda

Na arquitetura acima, fica nítida a divisão da SA em três etapas distintas (Sparck Jones, 1993): a versão do texto-fonte para uma forma de representação computável (representação conceitual I), a redução dessa primeira representação, dando origem ao sumário do texto-fonte, mas ainda sob a forma de uma representação não-textual (representação conceitual II), e, finalmente, a realização lingüística dessa representação (sumário resultante – abstract). É importante observar que as escolhas morfossintáticas, e mesmo as lexicais, não precisam coincidir com as feitas no texto-fonte. De acordo com o refinamento e com o grau de conhecimento lingüístico disponíveis no sistema, a tarefa como um todo se aproxima da sumarização humana, o que, potencialmente, pode nos aproximar de resultados mais satisfatórios.

Vale notar que o processo de SA lida com tipos distintos de informação (Rino & Pardo, 2003), uma vez que o próprio texto, por sua natureza, traz informações diferentes em seu bojo. Observam-se informações de cunho lingüístico, de cunho informativo (domínio) e de cunho comunicativo que se entrelaçam na construção da trama textual de acordo com padrões semânticos e pragmáticos, por isso a necessidade de modelos de representação do conhecimento capazes de lidar com esse grau de complexidade de informação.

2.3 Distinções entre os modelos de sumarização automática

Como se pode notar pela descrição das duas abordagens, o projeto e desenvolvimento de sistemas de Sumarização Automática são distintos: na profunda, o sumarizador automático deve incorporar modelos lingüísticos e/ou discursivos de interpretação e reescrita textual; na empírica, ele se baseia em modelos exatos de manipulação do conteúdo textual, sem que haja para todo o processamento interpretação ou estruturação do sumário pretendido. Diferentemente da profunda, a abordagem empírica prescinde da interpretação, baseando-se tão somente na distribuição de freqüência dos constituintes do texto-fonte a fim de elencar, para a formação do extrato, os mais significativos. Devido a essa diversidade, o processamento resultante caracteriza-se da seguinte forma:

- Abordagem profunda: alto grau de representação simbólica do conhecimento lingüístico e textual e raciocínio lógico baseado em técnicas simbólicas, para estruturação e reescrita do sumário;

- Abordagem empírica: processamento prioritariamente baseado em reconhecimento de padrões derivados de informações ou de suas distribuições numéricas. São usadas técnicas empíricas (modelos matemáticos) ou estatísticas para a extração dos segmentos textuais relevantes.

Como o foco deste trabalho encontra-se, mais particularmente, na abordagem profunda, e nosso objetivo é fornecer conhecimentos que minimizem problemas de textualidade causados por quebras de cadeias de co-referência, é necessário considerar os conceitos básicos de Lingüística Textual. Os principais são descritos no próximo capítulo.

3. Lingüística Textual: conceitos-chave para o tratamento computacional da textualidade

Nesta seção, apresentamos os conceitos fundamentais para uma abordagem científica dos textos considerados em sua qualidade textual. Tratamos da textualidade a partir do conceito de texto, coerência e coesão, aprofundando a investigação em um aspecto particular desta última: as cadeias de co-referência, particularmente os casos em que a expressão referencial é uma descrição definida (sintagma nominal iniciado por artigo definido). Interessa ao nosso estudo, sobretudo, a resolução anafórica dessas expressões. Esse foco restrito no estudo do fenômeno co-referencial deve-se, sobretudo, à necessidade de se abordar esta questão de um ponto de vista lingüístico-computacional, e não apenas lingüístico. O problema da perda da qualidade textual originada a partir das quebras de elos referenciais tem sido o enfoque de trabalhos tanto na área de SA (Seno, 2005; Seno & Rino, 2005) quanto na de resolução anafórica automática (Vieira, 2002; 2003).

3.1 Textualidade

Os sistemas de sumarização automática lidam com textos verbais na sua forma escrita. O estudo sistemático do texto é objeto da Lingüística Textual, uma ciência relativamente nova (os primeiros estudos específicos datam da década de 60) que, segundo Marcuschi (1983), funda-se em um princípio básico: o texto é uma unidade lingüística hierarquicamente superior à frase e a gramática da frase não é suficiente para descrever os fenômenos textuais.

Tal sistematização não deve ser confundida com o que se entende por “análise do texto” ou mesmo com a “análise literária”. O estudo feito na Lingüística Textual visa ao tratamento dos processos e regularidades segundo os quais se produz, constitui, compreende e descreve o fenômeno texto (Marcuschi, 1983). Desse modo, para o estudo da textualidade, é necessário partir do conceito de texto dado pela Lingüística Textual, sobre o qual, então, poderemos descrever quais atributos essenciais devem ser observados na estrutura textual e quais são os recursos lingüísticos que atuam nesse processo.

O texto, como veremos na seção 3.1.1, consiste em uma *realização verbal organizada*. Tal organização se dá tanto no plano da articulação semântica e pragmática (coerência textual), quanto no plano da estruturação interna dos constituintes do texto (coesão textual). Para a construção de um texto coeso, existem recursos de construção textual que vão desde as escolhas lexicais até a elaboração de cadeias de co-referência, sendo estas últimas o foco deste trabalho.

3.1.1 Noção de Texto

Um ponto fundamental para a elaboração de modelos que tenham por escopo um texto coerente e coeso é, exatamente, o que é um texto. Esta definição, porém, não é simples nem trivial como pode parecer. Ao longo da trajetória dos estudos daquilo que se pode chamar de Lingüística do Texto, ou Textual, observou-se a construção do *conceito de texto*.

Durante muito tempo, não houve uma preocupação pontual acerca da elaboração de um ramo da ciência que tivesse o texto como objeto de análise – o texto era um meio e não um fim. Foi só a partir de meados do século XX que se iniciou um movimento de elaboração de gramáticas textuais. Nessa fase, o texto era considerado apenas enquanto uma estrutura lingüística organizada – falando-se então em *texto* (constituintes lingüísticos *coerentemente* organizados) e em *não-texto* (constituintes lingüísticos organizados *sem coerência*). Segundo Koch (2004), nesta primeira fase, os conceitos de texto variaram desde "unidade lingüística (do sistema) superior à frase" até "complexo de proposições semânticas".

Observe-se que, nesse primeiro momento, o texto era considerado tão-somente como um produto, algo apenas analisado na sua interioridade. Não importava ao analista a vasta gama de elementos que tinham relação direta com a concepção do texto, o seu nascedouro. Mas essa visão mudaria consideravelmente com o nascimento e fortalecimento da Análise do Discurso, oficialmente “nascida” em 1969, com os trabalhos de Pêcheux, mas que ganhou força a partir das décadas de 70 e 80, principalmente com o trabalho de Michel Foucault.

Neste trabalho, é importante salientar, as denominações *texto* e *discurso* são tomadas como sinônimos ainda que, no contexto atual dos estudos lingüísticos, tal confusão já esteja

suficientemente elucidada. Nossas razões são de cunho essencialmente pragmático e levam em conta a cristalização do termo *discurso* nos estudos lingüístico-computacionais. Assim, cumpre estabelecer os parâmetros conceituais adotados neste trabalho, definindo não apenas *texto* (tendo *discurso* como sinônimo), mas também *gênero textual* e *tipo de texto*, que são conceitos utilizados de modos diferentes na Lingüística e na Lingüística Computacional.

Leontiev (*apud* Marcuschi, 1983) afirma que o texto não existe fora de sua produção ou de sua recepção. Essa idéia de levar em consideração o "em torno" (as condições de produção e recepção) permitiu uma maior flexibilidade com relação, principalmente, à separação entre o texto e o não-texto. Se, antes, texto era apenas uma construção organizada (coesa e coerente) que refletia uma competência cujos parâmetros estavam firmados na idealização da boa escrita, nesse novo momento o texto passou a ser considerado na sua totalidade, e muito do que antes seria considerado aberração ou sinal de incompetência, passou a ser analisado sob a ótica da intenção do produtor, objetivos de produção, alcance com relação ao receptor, estilo, gênero textual.

Gêneros diferentes de texto possuem prerrogativas distintas para que se separe o texto (produto textual) do não-texto (produto prejudicado pela má elaboração). Essa assunção é extremamente importante para várias pesquisas em Lingüística Computacional.

Segundo Bonini (2001), entre as abordagens mais conceituadas sobre gênero textual, na atualidade, estão as de Van Dijk (1979), Swales (1992), Biber (1988) e Bronckart (1987). À parte as peculiaridades, as quatro primeiras, embora relacionando ao contexto social de origem e associando propósitos comunicativos, concebem o gênero textual como uma estrutura composta de partes características, agrupadas sob determinada sintaxe que reproduz uma ordem canônica. Partem, no geral, de uma descrição do texto conforme é caracterizado e utilizado em determinado ambiente social, mas de forma generalizante e com a atenção voltada para os aspectos formais. Nesse sentido, De Beaugrande & Dressler (1981) afirmam que existem correspondências regulares entre a estrutura de um texto e a estrutura do mundo que o texto evoca. Esse conceito se aplica à estrutura geral que caracteriza a exteriorização do pensamento técnico-científico em sua forma mais abrangente.

Neste trabalho, adotamos o conceito de *gênero* proposto por Swales (1992, p. 58), que encerra as considerações feitas acima:

> Um gênero compreende uma classe de eventos comunicativos, cujos membros compartilham um conjunto de propósitos comunicativos. Estes propósitos são reconhecidos pelos membros especialistas da comunidade discursiva e, desse modo, constituem a justificativa do gênero. Esta justificativa dá forma à *estrutura esquemática do discurso* e influencia e restringe as escolhas de conteúdo e estilo. [grifo nosso]

Tomando os textos jornalísticos como um exemplo para análise, a partir da teoria de Swales é possível depreender não que o "texto jornalístico" é um gênero, e sim que o domínio jornalístico está constituído de gêneros (como notícia, reportagem, anúncio de classificados, artigo de opinião, charge etc.). Esses gêneros encontrados no domínio jornalístico é que são construídos a partir de "propósitos comunicativos" que dão forma à "estrutura esquemática do discurso". Afinal, os propósitos comunicativos e a estrutura esquemática de um artigo de opinião são completamente diferentes dos propósitos comunicativos e da estrutura esquemática de uma notícia (o único ponto em comum entre artigo de opinião e notícia é o fato de os dois serem publicados no meio jornalístico). Assim, artigo de opinião e notícia são gêneros textuais: o gênero textual é algo sempre bem específico, ligado aos gêneros do discurso ou da atividade (uma aula, por exemplo, ou uma venda por *telemarketing*, são gêneros do discurso).

Transportando o conceito acima para as discussões pertinentes à Lingüística Computacional, o gênero, considerado dentro de um domínio mais amplo (jornalístico, científico, publicitário etc.) pode ajudar a determinar as estratégias de processamento com as quais os sistemas de PLN irão lidar. Em vários aspectos, textos pertencentes a gêneros diferentes irão possuir características distintivas marcantes (o que Swales denomina “estruturas esquemáticas do discurso”) que podem ser fundamentais para a determinação das estratégias dos sistemas de processamento desses textos.

Tomemos o caso do analisador discursivo automático – DiZer (Pardo, 2005b) – que trata textos científicos, produzindo a estrutura retórica dos mesmos. No processamento feito por

esse sistema são consideradas as palavras e expressões indicativas (verbos que indicam relações ou frases cristalizadas como "o objetivo deste trabalho é") – elementos constitutivos do texto que, no processamento automático, são facilmente identificáveis e podem ser ponteiros (indicadores) para a interpretação automática. Todavia, para gêneros diferentes de texto esses elementos podem apontar para relações retóricas diferentes, prejudicando o desempenho do sistema.

Outro conceito relevante para os estudos delineados neste trabalho é o de *tipo de texto*. Leech (1983) afirma que existe uma certa vagueza quando se fala em organização do discurso (texto), principalmente porque não existe uma uniformidade no que se refere à definição de o que é uma *boa organização do discurso*. Segundo o autor, uma justificativa a essa dificuldade é a existência de diferentes modalidades de discurso, cada qual com sua maneira característica de organização interna. Desse modo, é preciso observar as dimensões em que os discursos diferem entre si, identificando as características que "tendem a permanecer estáveis em trechos razoavelmente longos", opondo a estas as "características que tendem a sofrer contínuas modificações durante o discurso" (Leech, 1983, p. 12). São essas características estáveis do discurso que nos permitem falar em *tipos de texto*.

Os dois conceitos apresentados anteriormente são similares e podem, facilmente, ser confundidos um com o outro. Assim, devemos compreender *domínio textual* como a estrutura esquemática mais ampla, característica de um discurso, ao passo que por *tipo de texto*, entendemos o conjunto de características que, pela continuidade e reiteração, permitem, dentro de um determinado domínio, subcategorizar os textos que, por sua vez, pertencem a *gêneros* distintos. Traduzindo esta distinção em um exemplo prático, tomemos o caso do texto jornalístico: o domínio é o que chamamos de *jornalístico*, mas dentro de um jornal há várias subcategorias de textos – comentário político, crítica de televisão, notícia policial etc. – que constituem, estas, os gêneros. Do mesmo modo, quando falamos em domínio científico, temos, para cada área do conhecimento, uma subcategoria, ou tipo de texto científico – o tipo de texto da Lingüística, o da Computação etc. – que podem pertencer a gêneros distintos: artigos científicos, *abstracts*, dissertações, monografias etc.

Em geral, apesar da possibilidade de especificação em três níveis (discurso>tipo>gênero), o mais comum é falarmos em domínio (mais amplo) e gênero (mais específico). A Figura 4 apresenta esquematicamente os níveis referidos.

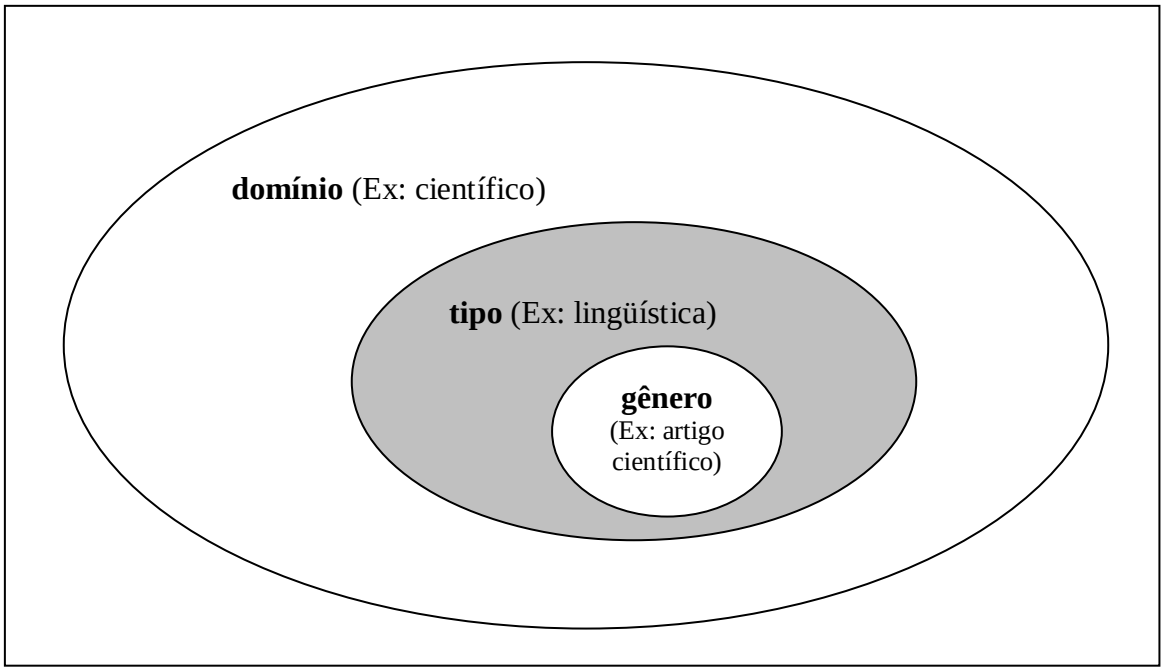


Figura 4. Representação gráfico-conceitual de domínio, tipo e gênero textuais

Em vista dessas considerações, optou-se, neste trabalho, por um conceito de texto mais amplo. Nos estudos da Lingüística Textual, de suma relevância é o trabalho de Koch (2004, p. 18), que nos propõe o seguinte conceito:

> "Poder-se-ia, assim, conceituar o texto como uma *manifestação verbal constituída de elementos lingüísticos selecionados e ordenados pelos falantes durante a atividade verbal*, de modo a permitir aos parceiros, na interação, não apenas a depreensão de conteúdos semânticos, em decorrência da ativação de processos e estratégias de ordem cognitiva, como também a interação (ou atuação) de acordo com práticas socioculturais (...) [O texto] deve preservar a *organização linear* que é o tratamento estritamente lingüístico, abordado no aspecto da coesão e, por outro lado, deve considerar a *organização reticulada ou tentacular*, não linear:

> portanto, dos níveis do sentido e intenções que realizam a coerência no aspecto semântico e funções pragmáticas". [grifos nossos]

O conceito acima pode ser traduzido na formulação genérica de que um texto possui, a priori, dois níveis de organização: a **coesão superficial**, que é a estruturação interna dos constituintes lingüísticos do texto; e a **coerência conceitual**, que abrange os aspectos semânticos e pragmáticos do texto. Esses dois níveis são indissociáveis no que tange à boa organização de um texto – a qualidade do texto depende igualmente dos dois – porém, é possível um estudo isolado de ambos (como veremos a seguir).

É senso comum na Lingüística Textual, hoje, que a textualidade é construída por um conjunto de fatores diversos, intra e extralingüísticos, que ainda escapa a uma enumeração precisa. De Beaugrande & Dressler (1981) propõem sete fatores fundamentais: coesão, coerência, informatividade, situacionalidade, intertextualidade, intencionalidade e aceitabilidade.

Dos fatores enumerados pelos referidos autores, apenas a coesão e a coerência estão adstritas aos limites intratextuais, razão pela qual são de maior interesse para sistemas que operam dados computáveis, como é o caso da Sumarização Automática.

É importante salientar que a Sumarização Automática contempla a informatividade, a intertextualidade (situacionalidade), a intencionalidade e a aceitabilidade. Por se tratar, porém, de uma abordagem computacional, a forma como esses fatores são tratados é diferente. A aceitabilidade, por exemplo, é medida tendo como restrição a aceitabilidade *do usuário* através de testes e avaliações.

3.1.2 Coerência textual

Segundo Koch (2004), a coerência se estabelece no nível semântico e cognitivo, estando relacionada ao sistema de pressuposições e implicações no nível pragmático da produção de sentido. Se, na coesão, fala-se em organização **linear**, a coerência é a organização **tentacular** (ou reticulada) – portanto, não linear – que articula os elementos lingüísticos

proeminentes na estrutura superficial e os níveis de sentido e intenções, conferindo ao texto plausibilidade semântica e pragmática.

De acordo com De Beaugrande & Dressler (1981), a coerência diz respeito ao modo como os componentes do universo textual, ou seja, os conceitos e relações subjacentes ao texto de superfície são mutuamente acessíveis e relevantes entre si, entrando numa configuração veiculadora de sentidos.

A coerência, pois, constrói-se em um nível que ultrapassa os traços lingüísticos do texto. A complexidade da coerência textual, em verdade, articula elementos de ordem lingüística, cognitiva e interacional – ou seja, não basta que haja conectividade lingüística entre os segmentos do texto, é preciso, mais, que haja relações de sentido entre os mesmos.

No que se refere ao trabalho ora exposto nesta dissertação, o tratamento da coerência limita-se às contribuições que a boa estruturação superficial pode fornecer à construção do sentido do texto. Os demais aspectos da coerência textual fogem ao escopo da abordagem de textualidade utilizada nos sistemas de processamento automático de língua natural que constituem o foco deste estudo.

3.1.3 Coesão textual

Na seção anterior, vimos que o texto, para ser coerente, precisa articular de maneira eficiente seus constituintes lingüísticos com a finalidade de alcançar o nível de sentido pretendido e, assim, promover a interação entre o texto e o leitor. Essa construção textual varia conforme o gênero e o tipo de texto, sendo que para alguns a organização linear é relevante e para outros não é.

A esta organização linear dos constituintes lingüísticos do texto dá-se o nome de *coesão*, que pode ser definida, sem muitas variações entre os autores, como sendo o uso de meios lingüísticos para facilitar a coerência (Halliday & Hasam, 1976; De Beaugrande e Dressler, 1981). Um elemento de coesão indica como a parte do texto na qual ela ocorre liga-se conceitualmente com uma outra parte no texto – a essa ligação dá-se o nome de *elos coesivos*.

Em determinados gêneros textuais – o lírico (poesia), em particular – a omissão de elos coesivos, ou mesmo estruturações pobres em elementos coesivos explícitos, constituem, não uma evidência de falha na estruturação, mas, antes, um poderoso recurso estilístico. Tomemos, a guisa de exemplo, um fragmento de "Alegria, alegria", de Caetano Veloso:

Caminhando contra o vento
Sem lenço sem documento
No sol de quase dezembro
Eu vou

O sol se reparte em crimes
Espaçonaves, guerrilhas
Em Cardinales bonitas
Eu vou

Em caras de presidentes
Em grandes beijos de amor
Em dentes pernas bandeiras
Bomba e Brigitte Bardot

O poeta não utiliza nenhum elo coesivo explícito na articulação das idéias expressas no texto. O encadeamento dos elementos que compõem a mensagem da canção (tipo de texto poético) dá-se não pela via estrutural, mas sim pelo ato de interação, ou seja, pelo próprio interlocutor no processo de comunicação (o leitor, no caso). A não-explicitação dos mecanismos coesivos representa, aqui, a aplicação artística da língua – a exploração de recursos figurativos.

Em domínios objetivos, como é o caso do jornalístico e do científico, as omissões de elementos estruturais que estabeleçam a coesão entre os elementos do texto, ou mesmo entre suas seções, constitui um *empobrecimento*. Cumpre ressaltar que, ao falarmos em empobrecimento do texto, não nos referimos à falta de textualidade, mas sim a 'déficits' de textualidade. Ao observarmos um sumário gerado automaticamente, podemos ter uma noção mais precisa do recorte que se pretende fazer:

> "O presidente da Venezuela, Hugo Chavez, pediu às famílias venezuelanas Em seu programa semanal de rádio transmitido no domingo, ele disse que o

Halloween é um jogo do terror, segundo a agência de notícias *Associated Press.*"[9]

A primeira sentença do sumário está incompleta em razão da falta do complemento verbal direto [o que o agente da ação pediu]. Porém, pela leitura do restante do texto, um leitor humano pode depreender – com certa segurança – que o pedido deve ser para que as famílias venezuelanas não participem/compactuem com o *Halloween*. Para efeitos de avaliação, porém, a má estruturação implica a deficiência do sumário no que se refere à textualidade.

Segundo Halliday & Hasan (1976), a coesão ocorre quando a interpretação de algum elemento no discurso é dependente da de outro. Um pressupõe o outro, no sentido de que não pode ser efetivamente decodificado a não ser por recurso ao outro. Trata-se de um conceito semântico de coesão, que explora as relações de sentido internas ao texto e responsáveis pela constituição do mesmo enquanto um texto. Assim, retomando o exemplo anterior, podemos identificar relações entre os constituintes lingüísticos do texto (elementos do discurso) que nos permitem, mesmo a partir do problema apontado (falta do argumento do verbo 'pedir'), inferir o sentido do texto.

Semanticamente, o texto destacado poderia ser expressado da seguinte maneira:

A (agente)
'pedir' (verbo – ação)
B (argumento indireto – destinatário do pedido)
C (argumento direto – o que é pedido)
D (justificativa do pedido)

A construção do tecido textual, segundo os autores, ocorre na medida em que vão se estabelecendo relações de sentido (semânticas) entre as sentenças encadeadas. A união

[9] Extraído do Córpus Rhetalho - http://www.icmc.usp.br/~taspardo/rhetalho.html.

dessas sentenças forma o que eles denominam de elo coesivo. Desse modo, consideremos o seguinte exemplo[10]:

[“O Supremo Tribunal Federal manifestou-se favoravelmente à pesquisa com células-tronco no Brasil.]$_1$ [***Com isso***, a perspectiva de investimentos em pesquisa pública e privada deve sofrer alterações drásticas”]$_2$

Nesse exemplo, a sentença (1) introduz uma informação que se relaciona com (2), sendo (2) uma decorrência lógica de (1) – nesse caso sendo possível a mesma interpretação ainda que não houvesse o marcador [com isso], que estabelece a relação de sentido entre as duas sentenças. Essa interpretação é possível porque o leitor utiliza seu conhecimento do mundo para compreender que a manifestação do STF é favorável ao desenvolvimento da pesquisa, ou seja, uma notícia favorável ao aumento dos investimentos na área de pesquisa. É, inclusive, essa a interpretação que podemos dar à expressão [alterações drásticas] – que poderia muito bem, em outro contexto, significar diminuição em lugar de aumento. Vejamos, porém, o mesmo texto de exemplo, apenas com uma alteração:

[“O Supremo Tribunal Federal manifestou-se favoravelmente à pesquisa com células-tronco no Brasil.]$_1$ [***No entanto***, a perspectiva de investimentos em pesquisa pública e privada deve sofrer alterações drásticas”]$_2$

A informação em (1) mantém-se inalterada, mas (2) sofreu uma modificação: o marcador, agora, não mais introduz uma relação de decorrência lógica, mas sim uma adversativa: (2) é uma situação em sentido contrário ao que se compreende em (1), e, assim, a interpretação que se dá a [alterações drásticas] deve acompanhar o sentido relacional introduzido pelo marcador e ser interpretado pelo leitor como uma diminuição.

Para Halliday & Hasan, portanto, a coesão e a coerência estão intimamente ligadas, uma vez que a segunda depende da adequação da primeira. Contrário a esse conceito de coesão textual, Marcuschi (1983) busca distinguir a coesão da coerência, afirmando serem conceitos separados. Segundo ele, é possível haver textos cuja organização linear se dê

[10] Texto criado para este exemplo.

apenas no nível do sentido, e não pela intermediação de constituintes lingüísticos. Um exemplo seria o seguinte texto[11]:

(a)

(1) O ódio no olhar. (2) Desejo de vingança. (3) Os pulsos que se contraem. (4) Fúria. (5) O prenúncio do soco. (6) Olhares se cruzam. (7) Confusão. (8) Sangue.

O encadeamento dos enunciados é totalmente desprovido de marcadores ou mesmo de uma estrutura mais elaborada. Todavia, pelo sentido dos enunciados e pela ordem de seu aparecimento, é possível vislumbrar um sentido, tornando o todo um texto. O encadeamento coesivo subjaz na progressão semântica dos substantivos que compõem o texto: ódio > vingança > pulsos contraídos > fúria > soco > olhar > confusão > sangue. Observemos que a ordenação destes itens lexicais obedece a uma construção progressiva do sentido, qual seja uma situação de violência.

Por outro lado, é possível haver uma seqüência de enunciados encadeados por recursos coesivos que, por falta de sentido, não formem um tecido textual:

(b)[12]

"Estava chovendo lá fora. Então o presidente caiu do palanque. Por causa disso o homem foi para a Lua e eu não entendi por que você roubou minha bicicleta."

O segmento acima possui uma estruturação coesiva evidenciada pela presença dos marcadores – então, por causa disso, e – mas, por faltar-lhe qualquer sentido lógico, ou seja, coerência, não se pode, em princípio, denominá-lo de texto[13].

O fato de ser possível a existência de textos desprovidos de elementos coesivos não é razão para a depreciação dos mesmos. Se, por um lado, existe a referida possibilidade – e tais textos pertencem a um domínio bastante restrito (literatura, texto telegráfico etc.) – a

[11] Texto criado para este exemplo.
[12] Texto criado para este exemplo.
[13] Ressalvada, aqui, a hipótese de se tratar de texto poético. A linguagem poética possui uma liberdade criativa cuja complexidade foge ao círculo no qual este trabalho está delimitado.

coesão confere ao texto maior legibilidade (Koch, 2004), sendo altamente desejável – quando não indispensável – na maior parte dos gêneros textuais (científico, didático, científico, jornalístico etc.). Isso porque a coesão explicita as relações existentes entre os constituintes do texto, possibilitando, assim, maior grau de compreensão e reduzindo as possibilidades de leituras equivocadas.

Observe-se que a coesão corresponde ao sentido lógico na estruturação dos constituintes lingüísticos do texto. Como observado no exemplo (a), a progressão semântica dos substantivos utilizados foi suficiente para que o sentido do texto (poético) fosse depreendido. O mesmo recurso é inadequado quando se tem, por exemplo, uma manchete jornalística, que utiliza outro tipo de estruturação. Em um caso e outro, porém, é evidente que a coesão é dada pelo encadeamento progressivo de itens lexicais nos quais se incluem os marcadores discursivos, mais abundantes em alguns gêneros e dispensáveis em outros.

Os recursos lingüísticos (mecanismos) disponíveis para a estruturação reticular do texto são variados. No exemplo (b), vimos o uso dos marcadores discursivos – *mas*, *porém*, *no entanto* etc.. Koch (2004) sistematiza o estudo destes mecanismos dividindo-os em duas modalidades: os de coesão seqüencial e os de coesão referencial, sendo esta última o foco deste estudo.

Koch (2004) define a *coesão referencial* como sendo aquela em que um componente da superfície do texto faz remissão a outro(s) elemento(s) nela presentes ou inferíveis a partir do universo textual. Nos estudos situados no campo da Lingüística Computacional, o termo comumente utilizado para se referir a esse fenômeno é **cadeia de co-referência**.

3.2 Coesão Referencial

Segundo Koch (2004), denomina-se **coesão referencial** "aquela em que um componente de superfície do texto faz remissão a outro(s) elemento(s) nela presentes ou inferíveis a partir do universo textual". A definição da autora engloba duas entidades referenciais distintas que devem ser identificadas e mais especificamente definidas a fim de que a distinção reste clara. Quando falamos em remissão de um componente do texto a outro nele presente, temos uma **forma referencial**, que também é chamada de **forma remissiva**. Isso porque,

no corpo do texto, é possível apontar uma entidade concreta como o termo antecedente à expressão referencial – a este termo antecedente dá-se o nome de **elemento de referência**, ou como Koch o denomina, um **referente textual.** O segundo caso descrito, quando um componente textual é compreensível por processos inferenciais, o elemento de referência (antecedente) não se encontra no texto, mas sim no contexto.

Esta distinção é importante porque nos permite compreender que, no processo de construção da referência textual, a ausência do elemento de referência no texto nem sempre implica a perda da referência.

Uma outra observação importante encontrada no trabalho de Koch (2004) é com relação à identidade entre o elemento de referência e a forma remissiva. Koch cita o trabalho de Kallmeyer et al. (1974) em que é feito um estudo aprofundado acerca da problemática da identidade referência-antecedente e no qual é proposta a **Teoria da Referência Mediatizada**, princípio que caracteriza a função mediadora exercida pela forma remissiva quando da remissão a outros constituintes do texto. A Figura 5 apresenta um modelo decorrente desta teoria (Kallmeyer et al., 1974 apud Koch, 2004).

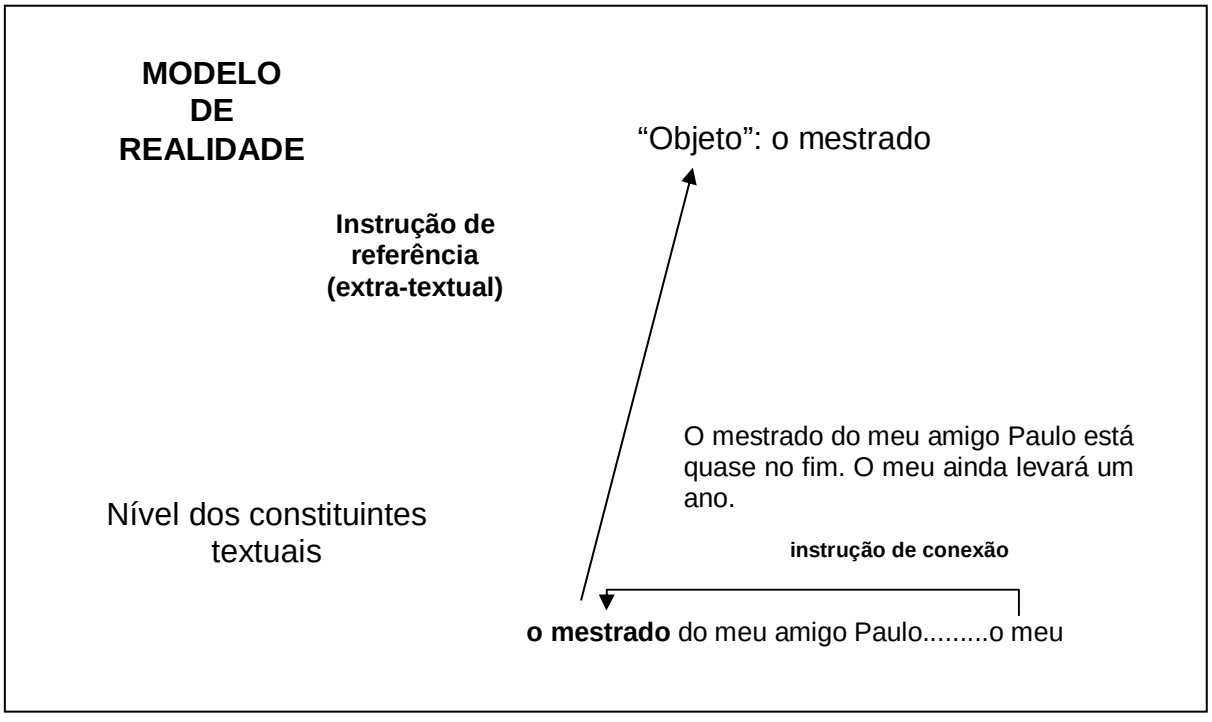


Figura 5. Modelo de representação da Teoria da Referência Mediatizada

Observemos que na frase ilustrada acima, o pronome possessivo [meu] não se refere ao grupo nominal [o mestrado de Paulo], mas apenas extrai desse sintagma o seu elemento de referência – *o mestrado* -, repudiando os demais elementos lingüísticos (amigo, Paulo, amigo Paulo).

A relação que existe em todos os casos em que um elemento lingüístico do texto, de alguma forma, remete a outro elemento introduzido no discurso previamente é uma relação de **encadeamento**, na qual é possível identificarmos o termo antecedente (referente) e a forma referencial (anáfora). Assim, no exemplo anterior, o encadeamento estabelece-se precisamente entre "mestrado" e "meu" e da compreensão desta relação depende a coerência textual do ponto de vista do leitor.

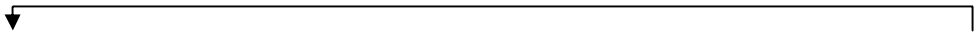
A esta construção dentro do texto, que estudamos mais detalhadamente na próxima seção, dá-se o nome de **cadeias de co-referência**.

3.3 Cadeias de co-referência

Um conceito que parece oportuno para nossa análise é o de cadeias de co-referência, que captura parte da "organização reticulada ou tentacular", mencionada por Koch (2004). Uma cadeia de co-referência num dado texto é o conjunto de todas as menções a uma determinada entidade/referente encontradas neste texto (Nenkova & Mckeown, 2003).

O fenômeno da co-referenciação manifesta-se quando da ocorrência de cadeias de co-referência em um texto, no qual o encadeamento entre referências textuais e referentes é parte do processo de estruturação coesa do texto. Segundo Mitkov (2002), a coesão é um atributo fundamental do texto; é o que separa o texto do amontoado de frases e orações sem conexão. O texto coeso é um conjunto de segmentos que mantêm uma relação entre si estabelecida, em geral, por entidades discursivas que possuem esta função específica.

Tomemos um exemplo[14]:



[14] Texto criado para este exemplo.

Margareth Thatcher comandou a Grã-Bretanha por anos a fio. A determinação *dela* era espantosa.

Neste segmento, é natural que liguemos o pronome [dela] ao nome [Margareth Thatcher], e dessa indexação depende a coesão do segmento. O termo [dela – de_ela] busca sua referência em uma menção anterior (daí a idéia de antecedente) e, ao estabelecer esta conexão, encontra sentido. No caso em que a forma referencial não identifica seu antecedente, a coesão textual resta comprometida e, neste caso, também a coerência. Vejamos o exemplo seguinte:

Luis Inácio Lula da Silva dirigiu-se ao país em um pronunciamento comovente. *A presidente* chegou a chorar.

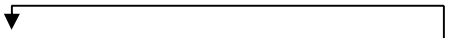
O sintagma nominal [a presidente] não consegue indexar nenhum elemento discursivo da oração anterior. Neste caso, portanto, ocorre a ruptura do elo referencial que, no texto, corresponde a um recurso de estruturação coesiva. Assim, tem-se um problema, causado pela impossibilidade da resolução anafórica entre o termo anafórico [a presidente] e seu referente, explicitado por um termo antecedente de maior representatividade semântica.

Esse problema, evidentemente, é de natureza **intratextual**, e representa uma limitação muito maior para a modelagem computacional que para a humana. Um interpretador humano (usuário) com um mínimo conhecimento de mundo para saber que [Luis Inácio Lula da Silva] corresponde a um homem, certamente superaria o engano patente no texto e perceberia que o SN [a presidente] indexa [Luis Inácio Lula da Silva], apesar da discrepância de gênero (Lula => masculino). Para a máquina, no entanto, a barreira é maior, uma vez que essa compreensão do engano ainda é uma limitação. Interpretadores humanos contam com o conhecimento pragmático para dar suporte à sua leitura e compreensão; interpretadores automáticos, não.

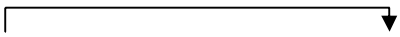
Para Halliday & Hasan (1976), a relação anafórica é o elo coesivo que remete uma palavra ao item anteriormente mencionado a ela correspondente. A palavra remetida é a anáfora; a palavra à qual se remete é o antecedente. E, por fim, o processo de determinação do termo antecedente para cada anáfora pode ser definido como resolução anafórica.

Como já visto, a construção de uma cadeia de co-referência é um processo que tem como pressuposto básico a existência de um ou mais termos (referências) que remetam a um outro termo, anteriormente mencionado no texto, com o qual se liguem. Mitkov (2002) indica a importância da distinção entre **antecedente** e **referente**. Para o autor, o referente é o ente pertencente ao mundo extralingüístico ao qual, tanto a referência quanto o termo antecedente remetem. O antecedente é, no mais, apenas a forma lingüística do termo que, na cadeia, é anteriormente mencionado e contém uma carga semântica mais completa.

A referência pode se estabelecer não apenas entre um termo posterior que faz remissão ao anterior (anáfora), como também é possível o oposto (catáfora). Observemos os exemplos:



George Bush visitou o Paquistão. Lá, *ele* foi recebido pelas autoridades do governo. [anáfora]



Ele foi um líder insensato, *o Bush*! [catáfora]

Nos diversos trabalhos que abordam cadeias de co-referência existentes na área da Lingüística Computacional, diversas são as classificações das formas referenciais (Vieira et al. 1998, 2003; Müller & Strube, 2001; Coelho et al., 2006). Em nosso trabalho, utilizamos o modelo de classificação proposto para os experimentos relativos ao Projeto ProCaCoSA, apresentados na seção 3.4.

3.4 Modelo de classificação de componentes de cadeias de co-referência

Inicialmente, é preciso esclarecer que neste trabalho de mestrado abordamos apenas uma fração do total de fenômenos co-referenciais. Nesta seção, descrevemos os parâmetros de anotação utilizados no projeto e todas as formas referenciais consideradas, estabelecendo, ao final, o recorte tratado nos experimentos pertinentes ao presente trabalho de mestrado.

O objetivo da anotação referencial no córpus Summ-it[15] (Collovini et al., 2007) abrangeu não apenas as relações anafóricas, mas também as referências dêiticas[16] realizadas por SNs, que podem ter como núcleo um nome ou um pronome, como podemos observar nos exemplos abaixo.

Ex: O presidente Lula, na última semana, assinou o acordo histórico que impulsionará a produção do biodiesel no Brasil. Lula afirmou que *o setor* deverá crescer mais de 20% apenas em 2007. *Ele* disse ainda que o álcool também receberá incentivos do governo.

A produção de biodiesel no Brasil > **o setor** (núcleo nominal)

O presidente Lula > Lula > **Ele** (núcleo pronominal)

Nas orientações de anotação, algumas decisões orientam os anotadores acerca do que não deve ser considerado, como é o caso das locuções articuladoras, tais como, "nesse sentido", "por essa razão", "como resultado", "além disso", "isto é".

A classificação proposta para a anotação de cadeias de co-referência leva em consideração a distinção entre SN com núcleo nome e pronomes. Compreendida a anotação como um processo, inicialmente são classificados os **sintagmas**, para, posteriormente, serem estabelecidas relações entre eles. As classificações abaixo correspondem aos sintagmas a serem identificados.

3.4.1 Sintagmas nominais com núcleo nominal

a) Sintagmas nominais formados por um núcleo do tipo "nome comum" (substantivo simples) antecedido por um determinante artigo definido – *o setor*, no exemplo abaixo.

Ex: A greve dos controladores de vôo paralisou os aeroportos do país. Os líderes d*o setor* exigem melhorias nas condições de trabalho.

---

[15] Na Parte II (aspectos práticos), reportamos a construção e as especificações técnicas do Córpus Summ-it. Resumidamente, trata-se do córpus construído para os trabalhos de pesquisa dos projetos ProCaCoSA e PLN-Br, composto por 50 textos jornalísticos da seção Ciência do jornal "A Folha de São Paulo" - http://inf.unisinos.br/~renata/laboratorio/desc_corpus_Summ-it.html

[16] Dêiticos são formas lingüísticas cuja referência só pode ser determinada pelo contexto.

b) Sintagmas nominais formados por um núcleo do tipo “nome próprio” (substantivo próprio) antecedido por um determinante artigo definido – *a Bolívia*, no exemplo abaixo.

Ex: O Estado Boliviano não pretende devolver as empresas nacionalizados em 2006 aos países de origem. *A Bolívia* não se ajoelhará mais, disse Evo Morales.

c) Sintagmas nominais formados por um núcleo nominal antecedido por em determinante artigo indefinido – *uma casa assim*, no exemplo abaixo.

Ex: A casa nova de Alice é um sonho. *Uma casa assim* deve ter custado milhões!

d) Sintagmas nominais formados por um núcleo nominal antecedido por um determinante pronome demonstrativo – *Esse traficante*, no exemplo abaixo.

Ex: Os policiais prenderam Fernandinho Beiramar no Morro do Cantagalo. *Esse traficante* é extremamente perigoso.

e) Sintagmas nominais formados por um núcleo nominal antecedido por um determinante pronome possessivo – *suas viaturas*, no exemplo abaixo.

Ex: A Polícia Militar está muito mal aparelhada. *Suas viaturas* estão em pandarecos.

f) Sintagmas nominais formados por um núcleo nominal antecedido por um determinante pronome interrogativo - *que nome,* no exemplo abaixo.

Ex: *Que nome* você escolheria para seu filho?

g) Sintagmas nominais formados por um núcleo nominal antecedido por um determinante numeral – *o primeiro trabalho*, no exemplo abaixo.

Ex: Existem vários trabalhos proposto nessa área. *O primeiro* foi o dos pesquisadores do Rio Grande do Sul.

h) Sintagmas nominais formados por um núcleo nominal antecedido por um determinante quantificador, incluídos aqui os pronomes indefinidos – *vários rapazes*, no exemplo abaixo.

Ex: Os estudantes chegaram em grande quantidade esse ano na cidade. *Vários rapazes* procuraram a pensão de Dona Eulália em busca de uma vaga.

i) Sintagmas nominais formados por núcleos nominais e coordenados entre si – *homens e mulheres*, no exemplo abaixo – *homens e mulheres*, no exemplo abaixo.

Ex: A grande massa humana lotou o estádio naquele domingo. *Homens e mulheres* acotovelavam-se nas arquibancadas lotadas.

j) Sintagmas nominais formados por um núcleo nominal não antecedido por qualquer determinante – *problemas dessa natureza*, no exemplo abaixo.

Ex: É horrível quando pelo contratempo de um assalto. *Problemas dessa natureza* são um transtorno.

k) Sintagmas nominais formados por um núcleo do tipo "nome próprio" não antecedido por qualquer determinante – *Lula*, no exemplo abaixo.

Ex: O presidente Lula viajou na última quinta-feira para o Chile. Em Santiago, *Lula* deu declarações à imprensa.

3.4.2 Pronomes

a) Sintagmas nominais formados por um pronome indefinido – *alguém*, no exemplo abaixo.

Ex: *Alguém* viu uma moça loira passar por aqui?

b) Sintagmas nominais formados somente por um pronome demonstrativo – *isso*, no exemplo abaixo.

Ex: O sujeito chegou bêbado em casa e espancou a esposa. *Isso* é um absurdo.

c) Sintagmas nominais somente por um pronome pessoal – *ele*, no exemplo abaixo.

Ex: Machado de Assis foi um gênio do Realismo Brasileiro. *Ele* é o autor de vários romances e contos.

d) Sintagmas nominais formados somente por pronomes possessivos – *minha*, no exemplo abaixo.

Ex: "De quem é esta caneta", perguntou o professor. "*Minha*", respondeu uma aluna.

e) Sintagmas nominais formados somente por um pronome interrogativo – *quando*, no exemplo abaixo.

Ex: Eu volto ao trabalho *quando* eles começarem a me pagar o que é justo.

f) Sintagmas nominais formados somente por um determinante numeral – *três*, no exemplo abaixo.

Ex: "Quantas garrafas de leite você quer?", perguntou a balconista. "*Três*, por favor", respondeu o cliente.

Com relação ao seu **estado** no discurso, os sintagmas podem ser classificados em:

a) **elementos novos no discurso:** o sintagma nominal introduz um novo referente no discurso sem apresentar parte de seu sentido ancorado em uma expressão anterior.

b) **elementos já mencionados no discurso:** o sintagma nominal é uma anáfora co-referencial, ou seja, retoma um referente já introduzido por uma expressão anterior.

c) **elementos associativos:** o sintagma nominal é uma expressão associativa, ou seja, introduz um novo referente no discurso cujo significado está ancorado em uma expressão anterior, mas com uma relação de dependência menor que aquela existente entre co-referentes.

d) **elementos dêiticos:** o sintagma nominal é uma referência dêitica.

No estudo do fenômeno co-referencial, o foco da análise recai sobre a relação observável entre os elementos novos no discurso e os já mencionados que lhes fazem referência. Pensando nessa relação, consideramos os SN anafóricos nos seguintes casos:

a) direto: a expressão anafórica e seu antecedente apresentam núcleos idênticos.

Ex: O notebook da Sony é o mais avançado no mercado. *Esse **notebook*** possui as seguintes funcionalidades (...)

b) indireto: a expressão anafórica e seu antecedente apresentam núcleos diferentes.

Ex: Paris é uma das capitais intelectuais do mundo ocidental. *A Cidade Luz* é o centro de onde emana a vanguarda artística de toda Europa.

c) encapsulation[17]: a expressão anafórica (inclusive pronomes) retoma um trecho de texto maior que um sintagma, tais como sentenças ou mesmo parágrafos.

Ex: Líderes do mundo inteiro reuniram-se na Suíça para discutir os níveis de emissão de carbono na atomosfera. (...) *O encontro* foi acompanhado de perto por ecologistas do Greenpeace.

Focalizamos, neste trabalho, as expressões referenciais do tipo **descrição definida**, ou seja, o SN antecedido por determinante artigo definido, restringindo nossa análise às referências diretas, indiretas e encapsulation, deixando de lado as do tipo associativo.

3.5 Descrições definidas

A discussão acerca da natureza e da função das descrições definidas no discurso humano é um assunto que extravasa o escopo de análise da própria Lingüística, penetrando o campo da Filosofia e da Lógica. As discussões levantadas nesses campos podem ser reduzidas às questões do modo como funcionam as descrições definidas de um ponto de vista lógico-semântico e como elas se relacionam com outras expressões, referencialmente. Interessa-nos, sobretudo, esta última.

Vieira (1998) apresenta uma vasta e completa revisão bibliográfica acerca das descrições definidas, tendo como foco a compreensão da função das mesmas no processamento do fenômeno co-referencial. Em seu amplo estudo do cânone dos autores mais significativos (Christopherson, Hawkin, Prince, Fraurud, Löbner, Clark, Sidner e Strand), Vieira verificou

[17] Optamos, neste trabalho, por manter a terminologia em língua inglesa a fim de permanecermos consistentes com relação ao Projeto ProCaCoSA, que decidiu pelo termo não traduzido.

que os diversos trabalhos listados indicam diferentes tipos de relações co-referenciais e associativas, sendo que cada uma delas pode ser realizada de diversas maneiras.

Segundo a autora, é possível dividirmos as descrições definidas em quatro classes, dependendo da forma com que estão relacionadas com os seus antecedentes. Se forem co-referentes, falamos em anáforas diretas e indiretas; se não forem co-referentes, falamos em expressões associativas (ou anáforas associativas) e em formas não anafóricas. A síntese desta classificação pode ser vista naFigura 6.

DESCRIÇÕES DEFINIDAS	**Co-referentes**	**Anáforas diretas**	Os pesquisadores encontraram-se n**o Congresso**. O balanço d*o Congresso* foi positivo.
		Anáforas indiretas	O corretor apresentou **o imóvel** aos interessados. Tratava-se d*o apartamento de um amigo*.
	Não Co-referentes	**Expressões associativas**	O Senador foi implicado n**o processo** pelo Conselho de Ética. *As investigações* continuam.
		Formas não anafóricas	Apresentamos **o novo grill de George Foreman**, perfeito para a culinária saudável. (entidade não mencionada anteriormente no discurso – discurso novo)

Figura 6. Classificação das descrições definidas

O nosso interesse neste trabalho é o estudo do fenômeno co-referencial envolvendo as descrições definidas em contexto co-referencial, tanto na forma de anáforas diretas, quanto de anáforas indiretas. Serão esses elementos lingüísticos que contemplaremos em nossa análise de córpus descrita nas seções adiante.

Considerações finais

Nesta seção vimos alguns conceitos lingüísticos básicos para a compreensão dos fenômenos tratados neste projeto. A partir da construção de um conceito de Lingüística Textual, exploramos a coerência e a coesão como atributos da textualidade e especificamos nosso recorte tomando por foco as cadeias de co-referência como elemento coesivo do texto. Apresentamos a classificação dos elementos co-referentes que orienta o estudo apresentado neste trabalho e determinamos nosso foco de estudo nas descrições definidas.

Na próxima seção, abordaremos as formas de tratamento computacional do texto, explorando, especificamente, as técnicas utilizadas para o tratamento computacional das

cadeias de co-referência, bem como a resolução anafórica em sistemas computacionais de PLN.

4. Modelos teóricos de representação do conhecimento lingüístico

Nesta seção, trataremos de alguns modelos teóricos que fornecem subsídios consistentes para a manipulação da informação textual, sendo, também, úteis em modelagens que privilegiam o tratamento das descrições definidas em cadeias de co-referência.

4.1 Teoria de Estruturação Retórica (*Rhetorical Structure Theory* – RST)

A RST é uma teoria relativamente nova que, apesar de desenvolvida para fins lingüísticos (Mann & Thompson, 1987), teve absorção expressiva na Lingüística Computacional, sobretudo para o processamento automático do inglês (Marcu, 1997a; Sporleder & Lascarides, 2005) e, mais recentemente, também do português (Pardo, 2005b; Seno, 2005). Todavia, trabalhos recentes (Desiderato Antônio, 2004) apontam para um resgate desse modelo teórico dentro da Lingüística, devido ao seu potencial de estruturação textual.

A RST fundamenta-se no princípio de que um texto tem uma estrutura retórica subjacente à estrutura superficial. Através dessa estrutura retórica, é possível recuperar o objetivo comunicativo que o escritor do texto pretendeu atingir ao escrevê-lo. Embora esses pressupostos da RST remetam ao aspecto discursivo, seu uso se limita à Lingüística Textual, ou seja, discurso somente enquanto texto. Este pode ser segmentado em unidades mínimas de significado (ou conteúdo) denominadas EDUs – *Elementary Discourse Units* – que, necessariamente, mantêm relação entre si na construção textual. Suas relações são previamente definidas pelo modelo, cujo conjunto, apesar de não definitivo, pretende ser suficientemente amplo para cobrir os casos retóricos considerados. Assim, as relações RST são divididas em duas classes: hipotáticas e paratáticas (Marcu, 1997a). As relações hipotáticas inter-relacionam pares de EDUs que apresentam diferentes graus de importância, sendo uma nuclear e a outra satélite. Essas relações denominam-se mononucleares. As relações paratáticas inter-relacionam EDUs que apresentam o mesmo grau de importância e são denominadas relações multinucleares.

Como forma de ilustrar os dois tipos de relação RST, vejamos os exemplos apresentados nas Figura 7 e Figura 8, para as sentenças (A) e (B) respectivamente. Os números entre colchetes nessas sentenças indicam a delimitação de cada EDU e os ramos em negrito nas estruturas RST, a informação nuclear (em oposição à satélite). Podemos observar em (A)

que a EDU 2 introduz o propósito (*purpose*) da EDU 1, sendo esta o núcleo e a outra, o satélite da relação PURPOSE. Em (B) temos uma relação de seqüência (*sequence*) entre os segmentos que não é hierárquica e, portanto, é multinuclear. As definições dessas relações na Teoria RST são apresentadas nas Figura 9 e Figura 10, respectivamente[18]. Optou-se, neste trabalho, por não traduzir os nomes das relações, haja vista a prática comum em outros trabalhos na área de Lingüística Computacional.

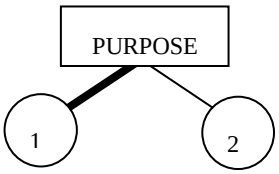
(A) O TSE enrijeceu as regras eleitorais validas já em 2006 [1] para evitar a prática de "recursos não capitalizados".[2]	

Figura 7. Sentença (A) e sua estrutura RST

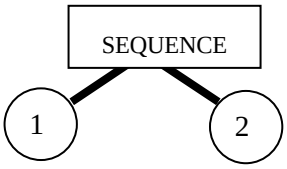
(B) Geraldo Alckmin visitou uma escola na periferia de São Paulo [1] e depois encontrou seus assessores no comitê central. [2]	

Figura 8. Sentença (B) e sua estrutura RST

Nome da relação: PURPOSE
Restrições sobre N: apresenta uma ação **Restrições sobre S**: apresenta uma situação não realizada **Restrições sobre N+S**: S apresenta uma situação que pode realizar N **Efeito**: o leitor reconhece que a atividade em N pode ser iniciada por meio de S

Figura 9. Definição da relação PURPOSE

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

Figura 10. Definição da relação SEQUENCE

[18] Definições extraídas de Pardo (2005); são traduções das originais em inglês (N: núcleo; S: satélite).

Como mostram os exemplos acima, a estruturação RST resulta em uma árvore e pode usar quaisquer relações do conjunto definido. Embora eles envolvam somente EDUs, isto é, unidades elementares do texto em foco, uma árvore RST é construída composicionalmente, ou seja, relações se estabelecem também entre subárvores RST, como mostra a estrutura RST para o Texto 1 exibida na Figura 11. Esse texto, extraído da Folha On-line (07/11/2006), foi analisado por um especialista em RST e sua estrutura foi construída manualmente, com o auxílio de uma ferramenta automática, a RSTTool (O´Donnel, 2000).

"[1]Os ministérios da Agricultura e da Ciência e Tecnologia defenderam ontem o uso da soja transgênica na produção do biodiesel [2] para abastecer parte da frota nacional de veículos.

[3] A idéia foi lançada pelo ministro Roberto Amaral [4] (Ciência e Tecnologia) [5] e detalhada ontem durante a abertura do 1º Congresso Internacional de Biodiesel, [6] realizado em Ribeirão Preto e [7] promovido pela USP [8] (Universidade de São Paulo) [9] da cidade. (...)"

Texto 1. Trecho do texto CIENCIA_2003_24219[19]

Na figura 11, podemos observar a estrutura arbórea composta por nós intermediários (relações) e nós finais (folhas ou EDUs), em cujas arestas estão indicados níveis hierárquicos de relacionamento das proposições (núcleo ou satélite).

[19] Fonte: Córpus Summ-it - vide texto completo no apêndice A.

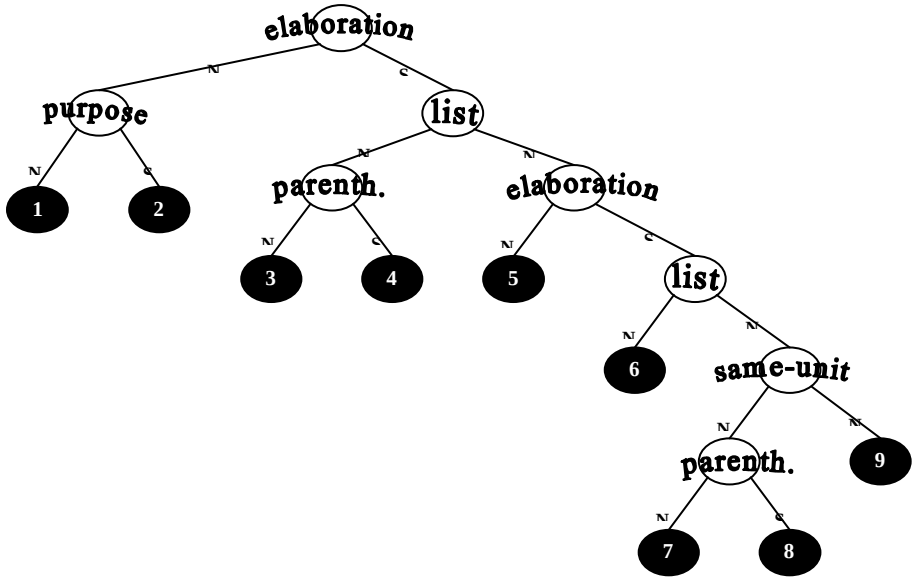


Figura 11. Estrutura RST do Texto 1

Inicialmente, os autores da teoria – Mann & Thompson (1987) – estabeleceram um conjunto de apenas 24 relações, as quais são exibidas na Figura 12, juntamente com seu tipo de nuclearidade. Outros autores, como Marcu (1997a), propõem conjuntos bastante diversos desse, remontando a mais de cem relações, o que torna um processo de análise bastante mais complexo. Como se pode notar na figura 11, algumas dessas relações têm significado evidente, já que são familiares a um falante competente, como a de propósito, evidência ou contraste. Entretanto, outras são mais obscuras, como ENABLEMENT e BACKGROUND. O leitor deve recorrer à obra de referência para entendê-las e recuperar seu contexto de uso.

Relação	**Mono-nuclear**	**multi-nuclear**	**Relação**	**Mono-nuclear**	**multi-nuclear**
ANTITHESIS	X		JUSTIFY	X	
BACKGROUND	X		MOTIVATION	X	
CIRCUMSTANCE	X		NON-VOLITIONA CAUSE	X	
CONCESSION	X		NON-VOLITIONAL RESULT	X	
CONDITION	X		OTHERWISE	X	
CONTRAST		X	PURPOSE	X	
ELABORATION	X		RESTATEMENT	X	
ENABLEMENT	X		SEQUENCE		X
EVALUATION	X		SOLUTIONHOOD	X	
EVIDENCE	X		SUMMARY	X	
INTERPRETATION	X		VOLITIONAL CAUSE	X	
JOINT		X	VOLITIONAL RESULT	X	

Figura 12. Conjunto original de relações RST

Para construir árvores RST como as ilustradas, o analista deve, primeiramente, reconhecer os tópicos principais de um texto e, assim, sua idéia principal, a fim de traçar o relacionamento retórico mais indicado entre os elementos macroestruturais. Entretanto, a construção da estrutura se dá primeiramente pela segmentação do texto em EDUs e pelo seu relacionamento, ou seja, pela construção de subárvores simples. Desse modo, tanto a recuperação da macro quanto a recuperação da microestrutura são relevantes. Ao mesmo tempo, dados os pressupostos da Teoria RST, a tarefa de análise visa recuperar as intenções do escritor. Assim, o analista deve ser, ao mesmo tempo, um leitor competente e um especialista na representação do conhecimento, para elaborar sua tarefa de modelagem artificial a fim de produzir uma estrutura RST segundo os pressupostos da Teoria. O conhecimento requerido envolverá padrões de análise, conhecimento morfossintático, reconhecimento de marcadores textuais e sua correspondência com as relações retóricas. Entretanto, a descoberta de padrões de análise é altamente dependente da apreensão da mensagem, ou idéia principal, do texto, assim como de sua organização macroestrutural. É por esse motivo que o conhecimento do próprio analista, usuário da teoria e leitor competente, se torna essencial. Em geral, os padrões indicativos das relações RST são dependentes de gênero e domínio textual: para sua determinação é preciso reconhecer a ordem das proposições (sejam elas EDUs ou segmentos textuais mais complexos), a qual será determinante da nuclearidade, isto é, do reconhecimento das unidades que serão núcleos e satélites, assim como de seu relacionamento funcional (determinação da relação retórica, propriamente dita).

A anotação[20] RST consiste, portanto, na recuperação do mapeamento de uma situação em língua natural, previamente elaborado pelo escritor. Nesse sentido, a situação espelha o modo e as razões de haver usos particulares da língua natural. Assim, tem-se por hipóteses que: i) a língua natural e a situação discursiva levam ao *efeito do discurso no leitor* – aqui é possível descobrir, por exemplo, por que os usos particulares da língua natural podem ter sucesso ou falhar ante o leitor; ii) o escritor deseja provocar, com seu texto, efeitos

[20] Trata-se do termo usual para o processo de análise com vistas à produção de um outro texto, o anotado com informações retóricas.

particulares no leitor; iii) o texto é fundamentado, portanto, nas intenções do escritor, conforme evidenciado pelas figuras 9 e 10.

O que dificulta a estruturação RST é a variedade de estruturas textuais e, certamente, a ambigüidade das intenções de um produtor e mesmo do receptor. Além disso, as próprias definições das relações RST são ambíguas, podendo levar a várias árvores RST para um mesmo texto. Para a estruturação, são considerados ainda os tipos básicos de estrutura textual, a saber: estrutura holística, relacional e sintática. A estrutura holística indica as propriedades do gênero ou variedade textual, que, por sua vez, remetem à macroestrutura textual; a relacional trata das estruturas linear (coesão) e reticulada (coerência) do texto. Por fim, a estrutura sintática simplesmente reflete a organização textual e discursiva. Claramente, esses tipos envolvem todas as condições a que se recorre intuitivamente para se recuperar o aspecto retórico de um texto.

Ao mesmo tempo em que a RST contempla essa variedade estrutural, ela também considera dois aspectos funcionais distintos ao definir suas relações: o apresentativo e o informativo (ou representacional). As relações funcionais são as relações de fato retóricas, ou discursivas, pois remetem aos efeitos que causam no leitor a partir dos objetivos do escritor, das suposições que ele faz sobre o leitor ou do uso de padrões específicos do domínio ou da audiência pressuposta. As apresentativas são as que servem de instrumento para a apresentação do assunto, ou conteúdo textual. Finalmente, as relações informativas espelham o meio (código ou instrumento) para transmitir a mensagem.

A questão da ambigüidade das relações está intimamente ligada à questão da subjetividade. O modelo, apesar da refinação na descrição das relações (para cada relação existe uma descrição minuciosa, com bastantes exemplos) não é suficientemente rígido e ainda ocorrem discrepâncias entre anotadores acerca de um mesmo texto (e mesmo discrepâncias de anotação do mesmo anotador que, ao revisar um texto já anotado por ele, anota-o diferentemente da primeira versão). Essa fragilidade, no entanto, pode ser minorada através do treinamento dos anotadores e da uniformização da anotação. Mann, Matthiessen & Thompson (1992), em trabalho posterior à formulação da RST, abordam pontualmente essa questão e indicam um conjunto de prerrogativas que devem orientar o analista RST, quais sejam: i) a idéia de **organização textual**; ii) a **unidade e a coerência** que existem entre as

partes do texto; iii) o fato de que essa unidade e coerência decorrem da **meta** do texto; iv) a **hierarquia** que se estabelece entre as partes do texto; v) **homogeneidade** da hierarquia; vi) **composição relacional;** vii) **assimetria das relações**; viii) **natureza das relações**; e ix) **número de relações**.

Os autores asseveram que, através de uma análise que contemple estes aspectos, é possível promover a uniformização da anotação, melhorando significativamente a qualidade dos trabalhos. Vejamos, então, estas premissas básicas com maior detalhamento que estão esquematizadas também na Figura 13:

1. **Organização textual**: um texto é constituído por partes funcionalmente significantes que se organizam com a finalidade de alcançar um determinado objetivo comunicativo. Ou seja, o texto possui uma organização que, à exceção das modalidades textuais artísticas (literatura e outros textos com pretensões literárias), é identificável para cada gênero e, dentro de cada gênero, para cada tipo de texto. Isso, por si só, é um ponto favorável à formalização;

2. **Unidade e coerência**: as partes do texto organizam-se de maneira a manter uma unidade, ou seja, se o texto tem um determinado objetivo discursivo, a unidade do texto é a organização das partes a fim de colaborar nessa meta. E para que tal se verifique, é preciso que as partes encadeiem-se coerentemente;

3. **A Unidade e a coerência decorrem da meta do texto**: basicamente, é o que foi explicado no item acima;

4. **Hierarquia**: as partes do texto dividem-se em porções mais relevantes e porções menos relevantes – é o que justifica a elaboração de heurísticas de poda que selecionem apenas seções relevantes do texto – o trabalho de Seno (2005);

5. **Homogeneidade da hierarquia**: dentro de uma estrutura relacional, a RST pressupõe a homogeneidade, ou seja, existe um conjunto de modelos de organização do texto que dão conta da análise desde da escala mais ampla até a mais refinada – estes modelos de organização do texto são os esquemas RST (*RST schemas*);

6. **Composição relacional**: há um conjunto restrito de relações que unem as partes do texto;

7. **Assimetria das relações**: a maior parte das relações é assimétrica, ou seja, entre as duas partes relacionadas existe uma hierarquia na qual uma pode ser definida como mais relevante que a outra (núcleo/satélite). Existem, também, em menor número, relações simétricas, chamadas de multinucleares;

8. **Natureza das relações**: fala-se em relações retóricas porque refletem as opções do autor do texto no momento da estruturação, organização e apresentação;

9. **Número de relações**: o conjunto de relações não é, ainda, fechado a novas relações que sejam identificadas pelos analistas. Assim, um analista pode acrescer relações que ache necessárias para descrever fenômenos encontrados em suas análises. Todavia, os autores afirmam que existe um conjunto restrito e ideal de relações universais que dará conta de descrever todos os fenômenos de estruturação retórica do texto e que, em princípio, não será aberto.

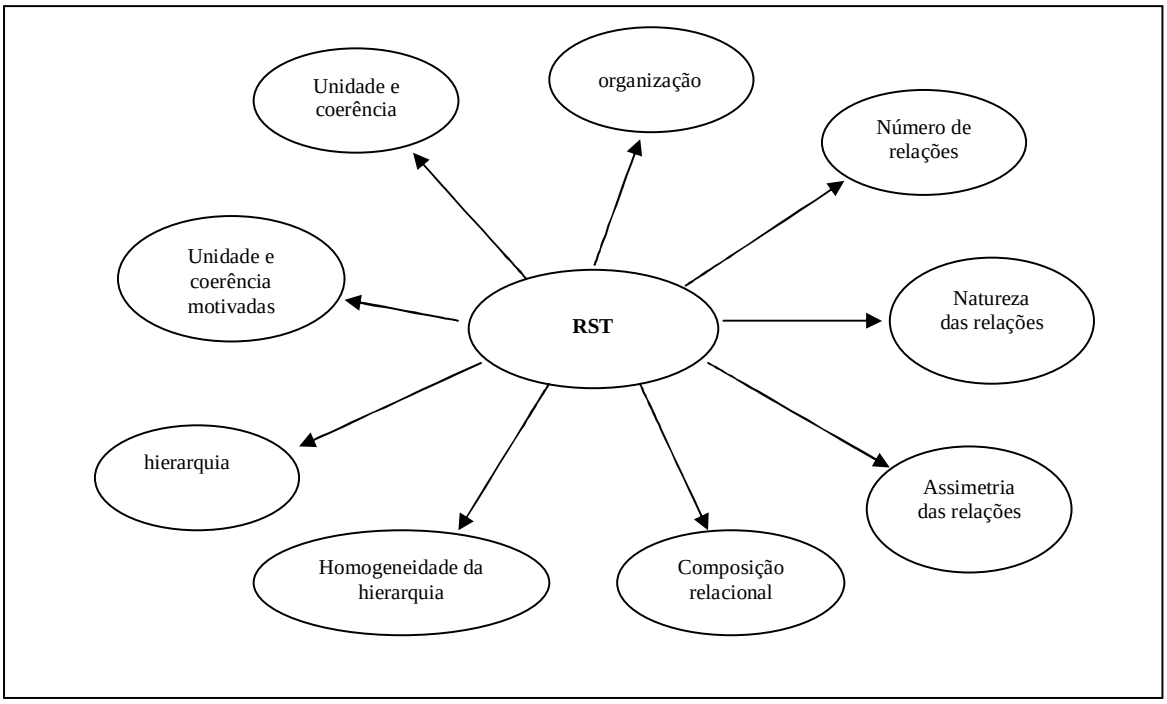


Figura 13. Prerrogativas básicas da RST (Mann, Matthiessen & Thompson, 1992)

Exemplo de análise RST manual

O exemplo a seguir foi retirado do trabalho de Mann, Matthiessen & Thompson (1992) e apresenta uma proposta de análise RST fornecida pelos próprios autores da teoria. O texto, simplesmente denominado de *ZPG Letter*, segue abaixo[21]:

{ZERO POPULATION GROWTH}$_{seg1}$
{November 22, 1985.}$_{seg2}$
{Dear Friend of ZPG:}$_{seg3}$

{At 7:00 a.m. on October 25, our phones started to ring.}$_{seg4}$ {Calls jammed our switchboard all day}$_{seg5}$. {Staffers stayed late into the night, answering questions and talking withreporters from newspapers, radio stations, wire services and TV stations in every part of the country.}$_{seg6}$

{When we released the results of ZPG's 1985 Urban Stress Test, we had no idea we'd getsuch an overwhelming response.}$_{seg7}$ {Media and public reaction has been nothing short of incredible!}$_{seg8}$

{At first, the deluge of calls came mostly from reporters eager to tell the public about Urban Stress Test results and from outraged public officials who were furious that we had "blown the whistle" on conditions in their cities.}$_{seg9}$

{Now we are hearing from concerned citizens in all parts of the country who want to know what they can do to hold local officials *acc*ountable for tackling population-related problems that threaten public health and well-being.}$_{seg10}$

{ZPG's 1985 Urban Stress Test, created after months of persistent and exhaustive research, is the nation's first survey of how population-linked pressures affect U.S. cities.}$_{seg11}$ {It ranks 184 urban areas on 11 different criteria ranging from crowding and birth rates to air quality and toxic wastes.}$_{seg12}$

{The Urban Stress Test translates complex, technical data into an easy-to-use action tool for concerned citizens, elected officials and opinion leaders.}$_{seg13}$ {But to use it well, we urgently need your help.}$_{seg14}$

{Our small staff is being swamped with requests for more information and our modest resources are being stretched to the limit.}$_{seg15}$

{Your support now is critical.}$_{seg16}$ {ZPG's 1985 Urban Stress Test may be our best opportunity ever to get the populationmessage heard.}$_{seg17}$

{With your contribution, ZPG can arm our growing network of local activists with the materials they need to warn community leaders about emerging population-linked stresses before they reach crisis stage.}$_{seg18}$

{Even though our national government continues to ignore the consequences of uncontrolled population growth, we can act to take positive action at the local level.}$_{seg19}$

{Every day decisions are being made by local officials in our communities that could drastically affect the quality of our lives.}$_{seg20}$ {To make sound choices in planning for people, both elected officials and the American public need the population-stress data revealed by our study.}$_{seg21}$

{Please make a special contribution to Zero Population Growth today.}$_{seg22}$ {Whatever you give - - $25, $50, $100 or as much as you can -- will be used immediately to put the Urban Stress Test in the hands of those who need it most.}$_{seg23}$

{Sincerely}$_{seg24}$
{Susan Weber}$_{seg25}$
{Executive Director}$_{seg26}$

---

[21] O texto encontra-se com a indicação da segmentação proposta pelos autores.

> {P.S.}seg27 {The result of ZPG´s 1985 Urban Stress Test were reported as a top news story by hundreds of newspapers and TV and radio stations from coast to coast.}$_{seg28}$ {I hope you´ll hepl us monitor this remarkable media coverage by completing the enclosed reply form.}$_{seg30}$

A análise feita pelos autores utiliza as vinte e quatro relações originais de Mann & Thompson (1987) (vide figura 12), e não o conjunto ampliado de Marcu (1999). O modo como foi procedida a análise também é relevante. A estruturação do discurso em unidades discursivas pode se dar de duas maneiras, basicamente: i) *bottom-up* (de baixo para cima): após a segmentação do texto (que pode ser oracional, sentencial, em parágrafos etc.), o analista observa primeiro as relações entre as unidades para, depois, contemplar relações entre blocos maiores de texto; ou ii) *top-down* (de cima para baixo): após estar o texto segmentado, o analista primeiro contempla as relações na macroestrutura para, posteriormente, refinar o estudo até a observância das relações entre as unidades mínimas (EDU). A abordagem dos autores foi *top-down*, iniciando-se pela investigação dos segmentos dotados de nuclearidade e, a partir desses segmentos, encontrando blocos que se relacionassem, chegando, ao fim, na inter-relação dos segmentos do texto.

Esse método de análise pode ser considerado mais preciso, pois identifica, liminarmente, o cerne do texto e, então, busca relações entre o restante do mesmo e essa porção nuclear. Desse modo, a visão da unidade e da coerência é mais translúcida, pois o analista, ao refinar sua análise (dissecando porções menores do texto e relacionando-as), já sabe qual é a meta a ser alcançada pelo texto, seu objetivo comunicativo.

No caso da carta ZPG, os autores já apontam como proposição central os segmentos nos quais a carta solicita a colaboração financeira dos leitores (segmentos 22 e 23). A partir dessa delimitação, o analista buscou blocos que se relacionassem de modo a colaborar com o sucesso desse fragmento nuclear. Nesse caso, um pedido de dinheiro precisava de uma motivação que convencesse o leitor a contribuir.

Assim, o corpo do texto em si foi identificado como uma longa construção textual com vistas a motivar o pedido de colaboração (segmentos de 4 a 21). Neste bloco, uma porção delimitada exerce papel nuclear (segmentos 11 a 16), pois são as alegações principais que

motivam o envio do dinheiro. Os blocos adjacentes, segmentos 7 a 10, e segmentos 17 a 21, exercem a função de evidenciadores de que aquelas alegações nucleares são consistentes.

Os autores dividem o texto em quatro blocos que são analisados separadamente: preliminares (título, cabeçalho etc), corpo da carta (possui estrutura relacional), encerramento (cumprimento de encerramento, assinatura, nomes etc.), e o *post scriptum* (que também possui estrutura relacional).

Identificado o objetivo comunicativo do texto – apresentar uma solicitação de contribuição financeira – os analistas identificaram também a porção nuclear do texto (segmento 22, com seu satélite 23). Essa forma de análise revela-se bastante interessante quando se leva em consideração que para determinados **gêneros textuais** é possível estabelecer-se a "zona provável" de nuclearidade. No caso em tela, tomada outra carta semelhante, solicitando contribuição, a construção desse tipo textual conduziria o analista, novamente, a buscar o núcleo comunicativo exatamente na solicitação – isso porque os textos que se inserem nessa categoria (quando bem escritos e estruturados) terão sempre como porção mais relevante o pedido dirigido ao leitor. Para outros gêneros textuais – tal como o texto jornalístico e o texto científico, por exemplo – também é possível identificar um padrão para o texto bem estruturado que permita a identificação da porção nuclear. Em textos jornalísticos curtos, tais como manchetes, a porção nuclear, com grande freqüência, encontra-se no primeiro parágrafo.

Os segmentos de 4 a 21 formam uma estrutura argumental que fornece subsídios para a motivação do pedido. Nesta porção de texto existe um núcleo também, identificado como as alegações principais que motivam a solicitação de dinheiro (segmentos 11-16). Adjacentes a esses segmentos, existem as evidências do núcleo, marcadas nos segmentos 7-10 e 17-21. Observemos na Figura 14 a estrutura arbórea da análise feita apenas do corpo da carta, desconsiderados os elementos textuais nos contornos do texto (título, data, assinatura etc.).

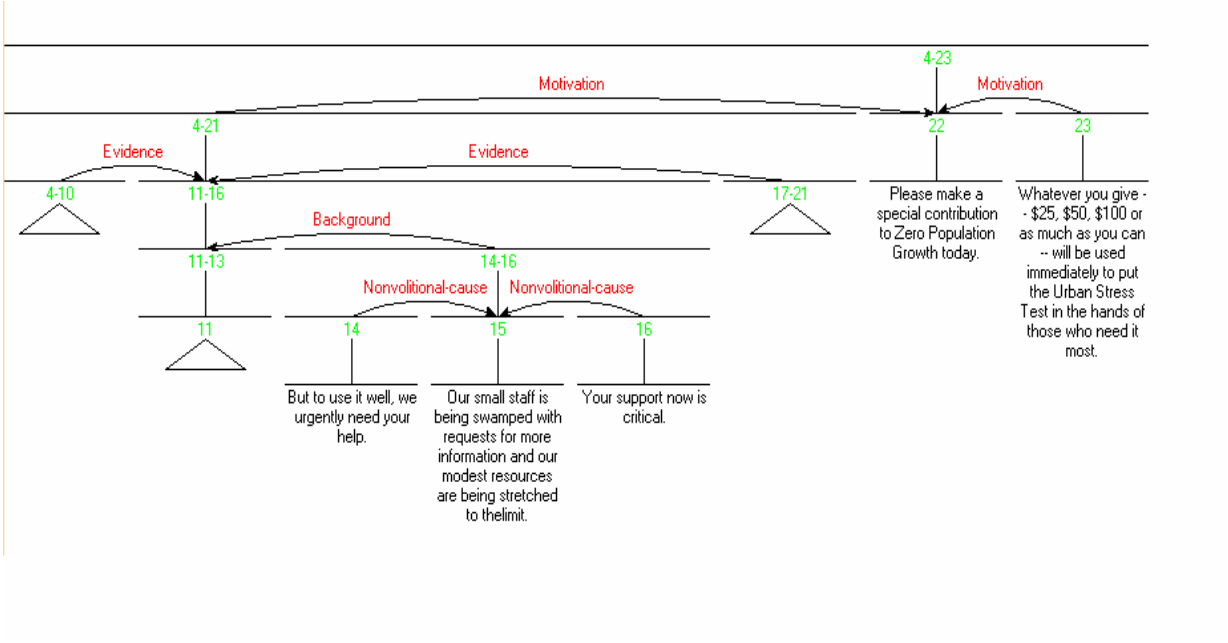


Figura 14. Estrutura arbórea da ZPG Letter

Podemos perceber pela representação gráfica da análise como os segmentos de texto apontam para o segmento 22, previamente identificado como núcleo do texto, ou meta.

Este tipo de representação textual, ou seja, o modelo RST, é de particular interesse ao trabalho desenvolvido neste projeto por se tratar da forma de representação do conhecimento adotada pelo RheSumaRST. Neste sistema, como veremos, o texto-fonte é na verdade representado por sua estrutura RST, que fornece a base para a aplicação do algoritmo de Cristea et al. (1998) (*Veins Theory* – Teoria das Veias, que permite a delimitação do domínio de acessibilidade referencial) e a posterior utilização de heurísticas de poda, gerando, assim, sumários.

4.2 Teoria de *Centering*

A Teoria de *Centering* (Grosz et al., 1995) parte do pressuposto de que um discurso deve apresentar coerência na seqüência de enunciados que o formam. Um discurso deve exibir **coerência global** – entre os seus diversos segmentos – e **coerência local** – entre as declarações de um mesmo segmento. A Teoria de *Centering* propõe um modelo para o tratamento da coerência local descrevendo um sistema de restrições e regras que governam as relações entre o **foco de atenção** do discurso e as formas escolhidas para construção das

declarações que o compõem. Exploraremos estes conceitos a partir de dois exemplos traduzidos de Grosz et al. (1995, p.203):

(A)[22]

(a) John foi à sua loja de música favorita para comprar um piano.

(b) Ele freqüentava a loja há vários anos.

(c) Estava excitado porque, finalmente, poderia comprar um piano.

(d) Ele chegou justo quando a loja fechava.

(B)[23]

(a) John foi à sua loja de música favorita para comprar um piano.

(b) Esta era a loja que John freqüentava há vários anos.

(c) Ele estava excitado porque, finalmente, poderia comprar um piano.

(d) Ela estava fechando quando John chegou.

Os dois segmentos de discurso expressam exatamente a mesma informação utilizando enunciados diferentes. No entanto, o discurso em (A) é intuitivamente mais coerente que em (B). Isso parece acontecer porque no primeiro caso trata-se apenas de um indivíduo central, John, enquanto que no segundo, o foco principal oscila entre John e a loja de música. Isto mostra que diferentes formas sintáticas implicam diferenças na inferência dos referentes anafóricos para o leitor (ou ouvinte). A teoria de *centering* fornece as bases para

---

[22] (a) John went to his favorite mosic store to buy a piano.
(b) He had frequented the store for many years.
(c) He was excited that he could finally buy a piano.
(d) He arrived just as the store was closing for the day.

[23] (a) John went to his favorite mosic store to buy a piano.
(b) It was a store John had frequented for many years.
(c) He was excited that he could finally buy a piano.
(d) It was closing just as John arrived.

tratamento destas diferenças. Isto foi usado por Brennan et al. (1987), para desenvolvimento de um algoritmo capaz de resolver anáforas.

Um segmento de discurso consiste de uma seqüência de enunciados U_1, U_2...U_n. Os enunciados possuem a propriedade de realizar entidades do contexto do discurso. Por exemplo, no enunciado [John foi a sua loja de música favorita comprar um piano], temos como entidades realizadas, [John], [loja-de-música] e [piano].

Os centros representam as entidades do mundo referidas pela sentença atual. Estes objetos servem de ligação entre um enunciado e outro no segmento de discurso que os contém. **Centros** são objetos semânticos, não são palavras, frases ou formas sintáticas. Além disso, o mesmo enunciado pode possuir centros diferentes em situações diferentes. Os centros de um enunciado podem ser classificados em: centros retrospectivos (*Backward-Looking Center)* e centros prospectivos (*Forward-Looking Centers*).

O centro retrospectivo (C_b) é uma entidade que estabelece uma ligação coerente com o enunciado prévio, sendo única para todo o enunciado que não seja o primeiro de um segmento. Supõe-se não existir esta espécie de centro relativo ao primeiro enunciado de um segmento, dado que tal enunciado não estabelece um vínculo com um anterior, já que pertence a um segmento diferente.

O centro prospectivo (C_f) se apresenta na forma de um conjunto. Constituem um conjunto de entidades ordenadas segundo algum critério de saliência, que fornecem possíveis ligações para o próximo enunciado. A entidade mais altamente classificada entre os centros prospectivos é o próximo centro preferencial (*preferred center*), doravante (C_p). Um critério usual para ordenação deste conjunto de entidades é a função gramatical: sujeito>objeto>... para X>Y, significando que X é o centro preferencial em relação a Y..

Abaixo, um exemplo que serve como ilustração dos termos introduzidos:

(C)

(a) John[a_1] possui um BMW[a_2] .

(b) Ele[r_1] dirige rápido.

(c) Pedro[a_3] corre com ele[r_2] no feriado[a_4] .

(d) Ele[r_3] freqüentemente o[r_4] vence.

Como pode ser observado no exemplo, as entidades do mundo são reconhecidas através dos substantivos que as descrevem: [John], [BMW], [Pedro] e [feriado]. Os índices que aparecem nos enunciados (a_1, a_2, a_3 e a_4) nomeiam as construções do segmento de discurso que referenciam as entidades. Da mesma forma, nomeiam-se os elementos anafóricos encontrados (r_1, r_2, r_3 e r_4). Com a classificação completada, pode-se construir a âncora, o par $\mathbf{C_b,[C_f]_i}$, de cada enunciado:

(C´)

(a) C_b(?), [C_f(joao, a_1), (bwm, a_2)]$_i$

(b) C_b(joao, a_1), [C_f (joao, r_1)]$_i$

(c) C_b(joao, r_1), [C_f(pedro, a_3), (joao, r_2), (feriado, a_4)]$_i$

(d) C_b(pedro, a_3), [C_f(pedro, r_3), (joao, r_4)]$_i$

O C_b apenas indica quem é a atual entidade central do discurso. No conjunto C_f, por sua vez, temos a relação entre as entidades e suas realizações. O par (pedro, r_3) da sentença (d), por exemplo, indica que a entidade [Pedro] foi realizada pelo elemento r_3. A primeira sentença apresenta um (?) como C_b. Isto acontece porque a teoria não define como escolher o C_b do primeiro enunciado do segmento de discurso.

A partir desses conceitos de “centro”, é possível verificarmos três tipos de relações válidas entre enunciados que são expressas pelas transições continuidade (*center continuation*), retenção (*center retaining*) e deslocamento (*center shifting*), caracterizadas como:

i) Continuidade: o C_b de um enunciado U_{n+1} é o mesmo que o C_b de U_n, e essa entidade é o elemento mais altamente classificado no C_f do enunciado U_{n+1};

ii) Retenção: o C_b do enunciado U_{n+1} é o mesmo que o C_b de U_n, mas essa entidade não é o elemento mais altamente classificado no C_f de U_{n+1};

iii) Deslocamento: o C_b de U_{n+1} é diferente do C_b de U_n.

Além dessas transições possíveis dos centros, a teoria propõe também duas regras, uma para realizações de centros correlatos e outra para transições preferenciais entre centros:

i) Regra 01: Se qualquer elemento de $Cf(U_n)$ é realizado por um pronome em U_{n+1}, então o $C_b(U_{n+1})$ precisa ser realizado por um pronome também.

ii) Regra 02: Seqüências de CONTINUE são preferidas a seqüências de RETAIN. Estas, por sua vez, são preferidas a seqüências de SHIFT.

A fim de ilustrar essas transições, tomemos oumexemplo de Grosz et al. (1995, p. 217):

(D)[24]

(a) John tem tido muitos problemas para organizar suas férias.

(b) Ele não consegue arranjar ninguém para assumir seus compromissos. [Ele => John]

(c) Ontem, ele telefonou a Mike para elaborar um plano. [Ele => John]

(d) Mike o tem aborrecido muito recentemente. [o => John]

(e) Ele ligou para John às 5:00 da manhã na sexta-feira, na semana passada. [Ele => Mike]

---

[24] (a) John has been having a lot of trouble arranging his vacation.
(b) He cannot find anyone to take over his responsabilities.
(c) He called up Mike yesterday to work out a plan.
(d) Mike has annoyed him a lot recently.
(e) He called John at 5 A.M. on Friday last week.

Podemos representar os movimentos de transição nos enunciados deste exemplo através da seguinte maneira:

(a) C_b(?); C_f (John, muitos problemas, suas férias); Transição => inexistente;

(b) C_b(John); C_f (John, alguém, seus compromissos); Transição => CONTINUE

(c) C_b(John); C_f(João, Mike, um plano); Transição => CONTINUE

(d) C_b(John); C_f(Mike, John); Transição => RETAIN

(e) C_b(Mike); C_f(Mike, John, 5 da manhã, sexta-feira da semana passada); Transição => SHIFT

O exemplo (D) ilustra as transições continuidade, retenção e deslocamento. Observe que os enunciados (a), (b) e (c) referem-se a [John] e [John], que é a entidade mais altamente classificada no conjunto C_f. Em (d) o centro C_b continua sendo [John], mas [John] já não é a entidade mais altamente classificada no conjunto C_f. Em (e),essa transição é efetivada pela transição do centro [Jonh] para o centro [Mike].

4.3 Teoria das Veias (Veins Theory – VT)

A Teoria das Veias (VT – Cristea et al., 1998) generaliza – ou melhor, "globaliza" – o conceito de coerência local proposto por Grosz et al. (1995) e parte de uma hipótese de análise retórico-discursiva nos moldes da Teoria de Estrutura Retórica (RST), desenvolvida por Mann & Thompson (1987).

A VT é construída sob a hipótese de uma variante da RST, explorando apenas a oposição hipotaxe-parataxe, ou seja, de forma totalmente ortogonal a qualquer tipologia de relações. Formalmente, a VT ignora rótulos de nó em favor dos de aresta. Para qualquer árvore discursiva *t*, a teoria postula uma função $acc_t : edus(t) \rightarrow 2^{edus(t)}$ – ou seja, um mapeamento de cada EDU de *t* para um subconjunto das EDUs de *t* – dita *domínio de acessibilidade referencial* e alega que referências feitas a partir de uma EDU *x* só podem ser resolvidas em

$acc_t(x)$[25]. Trata-se da *Conjectura C1* da VT, que pode ser mais bem detalhada como segue: para toda árvore *t*, toda EDU $x \in edus(t)$ e toda expressão referencial *e* ocorrendo em *x*, o que se denota por $e \in x$, uma das seguintes afirmações deve ser verdadeira:

caso I: ou *e* é *nova no discurso*, ou seja, realiza a primeira menção ao seu referente – é o caso da primeira menção a uma determinada entidade dentro do discurso – Ex: Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de [*uma nova espécie de dinossauro no Brasil*]$_{\text{nova no discurso}}$.

caso II: ou *e* é anafórica e a primeira menção ao seu referente é realizada em alguma EDU $y \in acc_t(x)$ - Ex: [O atacante Nelsinho sofreu uma contusão no jogo do último domingo]$_y$ (...) [*O jogador* deverá ficar fora dos campos por duas semanas]$_x$

caso III: ou *e* é anafórica e existem EDUs $y \in acc_t(x)$, $z \in acc_t(y)$ e expressões referenciais $e' \in y$ e $e'' \in z$ tais que *e*, *e'* e *e"* são co-referentes e *e"* é nova no discurso. Nesse caso, *e'* funciona como intermediário para que *e* acesse a menção inaugural *e"* - Ex: [Romário entro na "cruzada" do milésimo gol]$_z$ (...) [*O atacante do Vasco* afirmou que fazer o gol é uma questão de tempo]$_y$ (...) [*O jogador* saiu do Maracanã frustrado]$_x$;

caso IV: ou a referência em *e* pode ser compreendida na ausência das menções anteriores ao seu referente, como se fosse uma entidade inserida no discurso naquele momento. Os autores denominam esses casos de *referências inferenciais* – ocorre quando, apesar de uma expressão se ligar referencialmente a um antecedente, sua compreensão pode ser alcançada através de uma inferência do leitor, prescindindo, assim, do antecedente. É o que ocorre com as anáforas diretas (mesmo núcleo nominal que o antecedente) – Ex:

---

[25] O número de elementos do conjunto de partes de *S* é sempre maior que o número de elementos de *S*, mesmo no caso de *S* ter um número infinito de elementos.
Se *S* tem *n* elementos, pode-se provar que P(S) tem 2^n elementos. No caso de *S* ser um conjunto infinito, define-se $2^{|S|} = |P(S)|$ (em que |A| representa o número de elementos de A). Por outro lado, sendo $\aleph_0 = |\mathbb{N}|$, também pode ser provado que $2^{\aleph_0} = |\mathbb{R}|$.

["A estimativa anterior, da ONU (Organização das Nações Unidas), calculava (...)]$_1$ [segundo *a ONU*, o uso de mais de 40% das reservas"]$_2$ – nesse exemplo, o leitor não precisa do termo antecedente (1) para compreender o sentido da expressão anafórica (2), pois o núcleo nominal (ONU) é o mesmo e também porque a sigla "ONU" é de conhecimento geral. Em outros casos envolvendo siglas, o antecedente talvez não seja dispensável – [Os pesquisadores apresentaram seu trabalho no RANLP (Recent Advances in Natural Language Processing) (…)]$_1$ [O RANLP deste ano contou com a presença de Ruslan Mitkov entre outros pesquisadores.]$_2$ – nesse outro exemplo, a anáfora presente em (2) precisa remeter ao antecedente (1) para que o leitor entenda o que a sigla (pertencente a um domínio muito restrito) significa.

A definição de acc_t é relativamente simples, presta-se a uma implementação direta e pode ser encontrada no artigo original de Cristea et al. (1998) – corresponde ao conjunto de EDUs inseridas na veia que estão antepostas à EDU a qual pertence o *acc*. A Figura 16 apresenta graficamente acc_t, para *t* igual à árvore também representada na mesma figura. Nesse grafo de acessibilidade referencial, os vértices representam EDUs; e existe um arco de *x* para *y* se e somente se $y \in acc_t(x)$, ou seja, se a EDU *y* é diretamente *acc*essível a partir de *x*.

Consideremos o seguinte fragmento de texto[26] na Figura 15:

[Os ministérios da Agricultura e da Ciência e Tecnologia defenderam ontem o uso da soja transgênica na produção do biodiesel]$_1$ [para abastecer parte da frota nacional de veículos.]$_2$
[A idéia foi lançada pelo ministro Roberto Amaral]$_3$ [(Ciência e Tecnologia)]$_4$ [e detalhada ontem durante a abertura do 1º Congresso Internacional de Biodiesel,]$_5$ [realizado em Ribeirão Preto]$_6$ [e promovido pela USP]$_7$ [(Universidade de São Paulo)]$_8$ [da cidade.]$_9$

Figura 15. Fragmento do texto CIENCIA_2003_24219

[26] Texto CIENCIA_2003_24219 do córpus Summ-it (vide apêndice A).

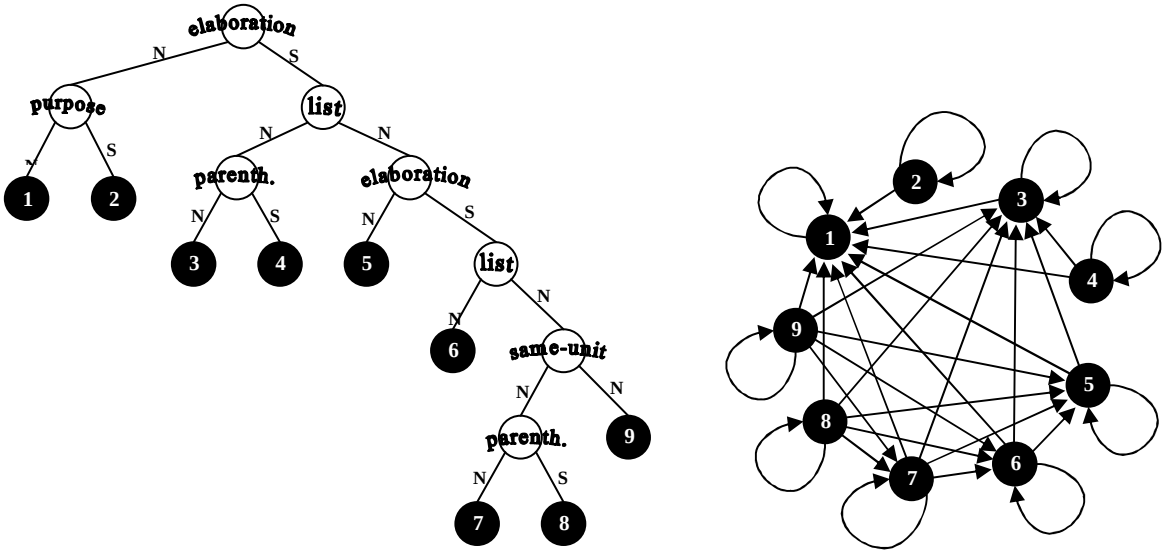


Figura 16. Árvore RST e Grafo de acessibilidade referencial.

Segundo a VT, as veias definidas sobre uma árvore RST são subseqüências da seqüência de unidades elementares que compõem o discurso e são determinadas de acordo com o seguinte algoritmo (Cristea et al. 1998, *apud* Seno, 2005):

Para todo n $\in$ *ARST*

Se n é um nó folha

então head de n é igual a n

Senão

head de n é igual à concatenação das heads dos seus filhos nucleares

Se n é o núcleo raiz da ARST, isto é, o núcleo mais nuclear

então veia de n é igual a sua head

Para todo n núcleo cujo n pai tem uma veia v

Se n tem um irmão satélite à sua esquerda com head h

então veia de n é igual a seq(mark(h), v)

Senão

veia de n é igual a v

Para todo n satélite de head h cujo n pai tem uma veia v

Se n é o filho esquerdo do seu n pai

então veia de n é igual a seq(h,v)

Senão

veia de n é igual a seq(h, simpl(v))

para:

ARST: árvore RST de um texto-fonte qualquer;

n: nó da ARST em foco;

head de n: conjunto de unidades mais salientes de n, isto é, as unidades mais importantes no segmento de discurso correspondente;

mark(x): função que dada uma string de símbolos x, retorna cada símbolo em x marcado de alguma forma (por exemplo, com parênteses ou colchetes);

simpl(x): função que elimina todos os símbolos marcados dos seus argumentos (se existir algum), por exemplo, simpl(a(bc)d(e)) retorna ad;

seq(x, y): função que pega como entrada duas strings não-intersectadas de nós folhas, x e y, e retorna a permutação de x concatenado a y, dada pela seqüência de leitura de x e y na ARST.

A mesma informação, representada por esse grafo na figura acima, pode ser representada através de uma árvore com as informações (veias e *acc*) indicadas por etiquetas em cada EDU, como fica explicitado na Figura 17.

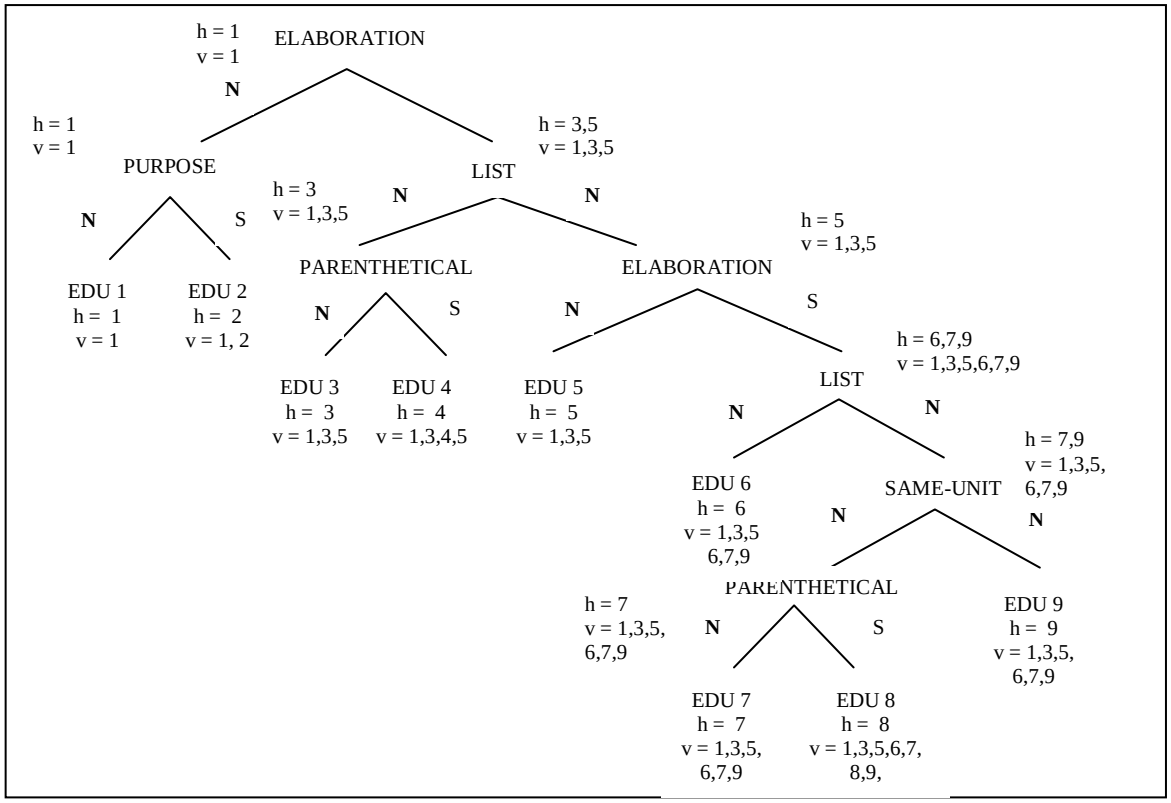


Figura 17. Representação das veias na forma de etiquetas na estrutura RST

No texto usado como exemplo, a EDU 9 “da cidade”, remete diretamente ao antecedente “Ribeirão Preto”, inserido na EDU 6. Se observarmos na Figura 17, a veia da EDU 9 contém a EDU 6, comprovando, assim, o poder preditivo proposto pela teoria.

É importante observar que, apesar de já utilizada por sistemas desenvolvidos no Brasil e que manipulam textos em português, a Teoria das Veias não tinha sido, até agora, validada para a Língua Portuguesa. Os dados relativos ao cálculo de precisão existiam apenas para o inglês, o romeno e o francês e são reportados por Cristea et al. (1998). Neste trabalho, os autores apresentam dados realmente animadores: precisão quase absoluta para o inglês e totalmente absoluta (100%) para o francês e o romeno. A utilização da VT por sistemas para o português, como veremos adiante, apresentou resultados interessantes, porém não absolutos (nem quase). Por esse motivo, faz parte do escopo deste trabalho de mestrado uma validação desta teoria para a Língua Portuguesa, utilizando uma metodologia consistente e bem documentada – é o que veremos na seção 9.

.

PARTE II – ASPECTOS PRÁTICOS DO TRABALHO

O escopo do presente projeto de mestrado foi a investigação sistemática do fenômeno da co-referenciação em sumários automaticamente produzidos. Conforme visto nas seções prévias, existem sistemas computacionais atualmente disponíveis, e que conciliam variadas teorias lingüísticas com vistas ao tratamento das cadeias de co-referência no processamento de línguas naturais.

A proposta de pesquisa foi, portanto, deslindar a aplicação destas teorias nos referidos sistemas, apontando e classificando os problemas de resolução anafórica, identificando suas origens na estrutura do processamento computacional, visando novas abordagens que conduzam a melhores resultados no que se refere à qualidade textual dos sumários automaticamente produzidos.

Os problemas com os quais lidamos neste projeto de mestrado devem ser compreendidos dentro do contexto da automação do processo de SA. Inicialmente, partimos do modelo proposto por Seno (2005). Sua proposta de sumarizador automático, o RheSumaRST, agrega as teorias descritas na seção anterior (RST e VT), com o objetivo de manutenção dos elos co-referenciais. Todavia, apesar da aplicação da VT e da utilização de estruturas retóricas manuais (anotadores humanos), ainda persistem resultados que evidenciam quebras de cadeias de co-referência nos sumários gerados.

Nossa idéia, portanto, foi a de investigar, paralelamente, tanto a validade da VT para o português, quanto a estruturação retórica. Nossas hipóteses, nesse caso, eram: i) a precisão reportada pelos autores (Cristea et al.) para as línguas de avaliação da teoria não eram, de fato, realistas (ao menos para a Língua Portuguesa); ii) problemas na estruturação retórica (tanto manual quanto automática) levavam às referidas quebras.

Neste cenário, incluímos ainda um gerador automático de estruturas RST, o DiZer (Pardo, 2005), que, acoplado ao RheSumaRST, originou o RheSuma-2 (Carbonel et al., 2006). Dentre os experimentos que descrevemos neste trabalho, reportamos o referido acoplamento e avaliamos as perdas de referência que a imprecisão da geração automática

das estruturas retóricas reflete nos sumários automáticos em termos de textualidade, particularmente no tocante às cadeias de co-referência. A síntese do referido processo pode ser observada na Figura 18.

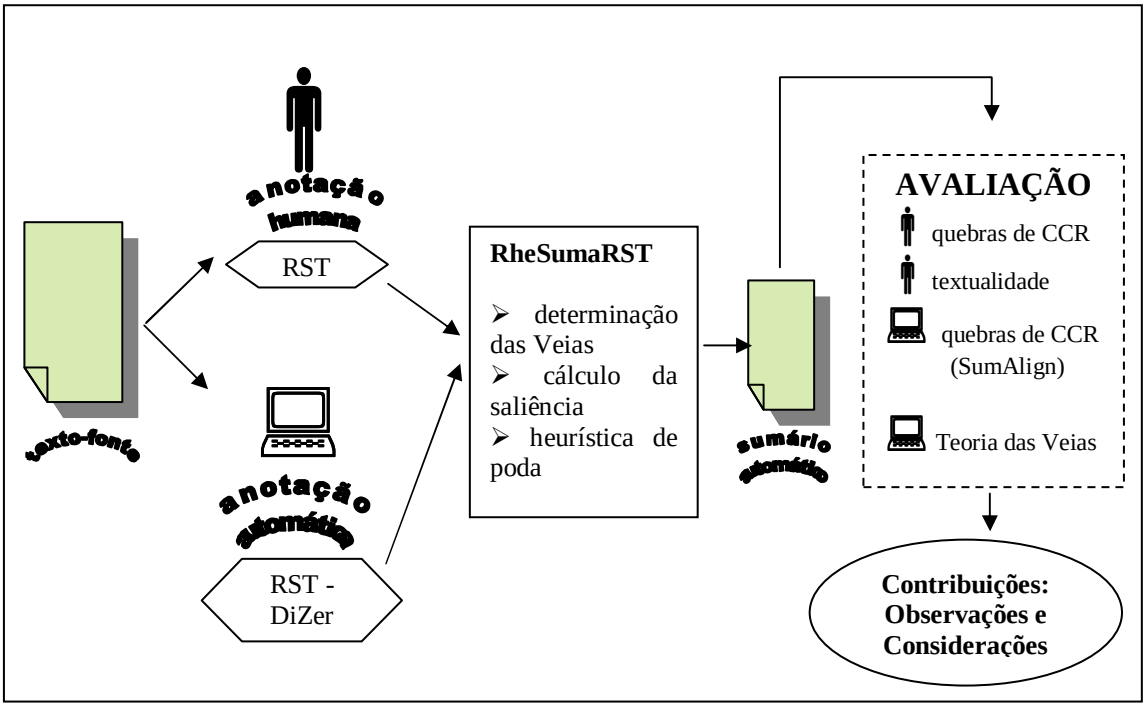


Figura 18. Cenário do projeto de mestrado

Neste preâmbulo à parte II, definimos, pontualmente, os objetivos de projeto, bem como as premissas e hipóteses consideradas na pesquisa e a motivação do projeto. Apresentamos o contexto dos experimentos já realizados, apontando as potenciais contribuições da pesquisa.

Objetivos

O objetivo central do projeto de pesquisa foi a propositura de métodos ou modelos visando melhorar o desempenho de sistemas de SA no que tange à qualidade textual dos sumários gerados. Para tanto, fez-se o estudo detalhado da co-referenciação enquanto um recurso para a construção da textualidade, tendo como foco a resolução anafórica em sistemas de sumarização automática. Este estudo incluiu, obviamente, não apenas o fenômeno lingüístico, mas também os sistemas computacionais, as formas de representação do

conhecimento utilizadas por tais sistemas e os métodos computacionais de processamento da língua com vistas à manutenção das cadeias de co-referência.

Nesse diapasão, temos a inter-relação de conhecimento lingüístico e perspectivas lingüístico-computacionais na persecução da melhora de sistemas de sumarização automática. Desvinculado de problemas de construção de sistemas computacionais, o conhecimento lingüístico pode ser tratado e fundamentado subjetiva e cognitivamente. Porém, para a elaboração computacional, é imprescindível que os aspectos lingüísticos sejam clara e inequivocamente formalizados. Logo, a metodologia de exploração de fenômenos lingüísticos não foi desvinculada de sua modelagem computacional. É por essa razão que foram sugeridos modelos de estruturação discursiva amplamente utilizados na Lingüística Computacional – RST e Teoria das Veias – e não somente modelos clássicos, teóricos, da Lingüística Textual. Também foram considerados sistemas reais, já em operação, para o estudo do fenômeno de co-referenciação vinculado às restrições de sumarização automática.

Sintetizando, portanto, apontam-se os seguintes objetivos que foram perseguidos no presente projeto de mestrado:

- Objetivo 01 (O1): a detecção de problemas de coerência decorrentes da ausência de resolução anafórica.

- Objetivo 02 (O2): a busca por um modelo de correspondência entre cadeias referenciais e estruturas retóricas, com foco especial em problemas de estruturação de sumários.

- Objetivo 03 (O3): propositura de meios de se relacionar e manipular a informação textual, a fim de averiguar como melhorar o desempenho de sumarizadores em relação à coerência dos sumários gerados.

Partindo da determinação dos objetivos, vejamos os passos para a consecução de cada um:

- Para O1:
  - Anotação automática de cadeias de co-referência em textos-fonte, com a ferramenta MMAX (Müller e Strube, 2001);
  - Verificação de suas cadeias de co-referência nos respectivos sumários ou extratos automáticos, para
    - Identificar e diagnosticar os problemas de co-referência;
    - Verificar a relevância das cadeias de co-referência para a sumarização automática;
    - Propor estratégias de identificação de problemas de referência.
- Para O2:
  - Anotação retórica de córpus de textos-fonte, isto é, geração de suas estruturas de discurso, ou estruturas RST, com a RSTTool (O'Donnell, 1997;
  - Demarcação das veias das estruturas RST anotadas, com ferramentas baseadas na RST e na Teoria das Veias (Cristea et al, 1998, Cristea, 2003), em construção pela equipe;
  - Análise da correspondência entre as cadeias de co-referência e as relações RST, considerando concomitantemente as veias do discurso-fonte;
  - Definição de modelos de sumarização de estruturas RST que possam identificar quebras de cadeias co-referenciais, visando à preservação da textualidade, concomitantemente com a crítica do modelo heurístico já existente.

- Para O3:

    - Indicação de possíveis elos coesivos entre informações morfológicas, lexicais, sintáticas ou semânticas, visando a incorporação desse conhecimento aos métodos extrativos.

    - Fazendo uso das tarefas de anotação de córpus de textos e sumários já discriminadas, apontar possíveis elos coesivos entre informações morfológicas, lexicais, sintáticas ou semânticas, visando à incorporação desse conhecimento aos métodos extrativos.

Premissas e hipóteses

Orientamos o presente trabalho de pesquisa segundo as seguintes **premissas**:

a) Um texto coeso e coerente pode ser submetido a um processamento do qual deve resultar um sumário com os mesmos atributos de textualidade;

b) A manutenção dos elos referenciais é necessária para que o um texto seja coeso, garantindo-se, assim, sua inteligibilidade. Um sumário, enquanto texto, deve seguir o mesmo princípio;

c) Para os gêneros textuais analisados neste projeto, a textualidade depende da coesão referencial.

As **hipóteses** estabelecidas foram:

a) Existe uma representação formal do texto que permite a sistemas de SA identificar informações que garantam o encadeamento referencial em sumários produzidos automaticamente;

b) É possível agregar a RST a este modelo de representação formal do texto – conforme sinaliza a Teoria das Veias;

c) Através de um estudo de córpus, é possível determinar estratégias de construção textual, depreendendo das mesmas informações para a construção do referido modelo de representação formal.

5. Córpus de trabalho: construção, anotação e análise

A metodologia deste projeto de mestrado é baseada, principalmente, na utilização de córpus para análise de fenômenos relacionados ao foco da pesquisa: estruturação retórica e referenciação.

Para tanto, foi necessária a construção de um córpus de textos anotados, tanto com cadeias de co-referência, quanto com estruturas retóricas (RST), a fim de que fosse possível: i)a validação das propostas de engenharia do conhecimento levantadas neste trabalho; ii) a validação do desempenho do modelo de SA em foco; e iii) a o estudo e validação de teorias do domínio lingüístico-computacional com vistas à melhoria do tratamento de cadeias de co-referência em sistemas de SA.

Nesta seção, descrevemos o processo de construção do córpus Summ-it[27] (Collovini et al., 2007), anotado com cadeias de co-referência (5.1) e estruturas retóricas (5.2).

5.1 Anotação de córpus com cadeias de co-referência

Córpus anotados manualmente podem ser utilizados tanto em etapas do processamento automático de alguns sistemas, como em tarefas de verificação de desempenho de outros sistemas automáticos, tais como sistemas de resolução de co-referência automáticos (Müller et al., 2002; Ng & Cardie, 2002; Poesio et al.; 2005), sistemas de sumarização automática (Seno, 2005; Seno & Rino, 2005; Carbonel et al., 2006), ou mesmo no sistema de verificação do algoritmo da Teoria das Veias, apresentado neste trabalho: o VeinTracker (Carbonel et al., 2007).

Para o inglês existem os córpus MUC-6 e ACE, disponibilizados pelo LDC (*Linguistic Data Consortium*)[28]. No contexto da Língua Portuguesa, o Córpus Summ-it, cuja construção reportamos nesta seção, configura-se como o pioneiro. Este córpus é formado por 50 textos jornalísticos retirados do caderno de Ciências da Folha de São Paulo, escritos

[27] http://inf.unisinos.br/~renata/laboratorio/desc_corpus_Summ-it.html
[28] http://www.ldc.upenn.edu/

em português do Brasil[29] e foi construído no âmbito do Projeto PLN-BR (Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil)[30].

A anotação de co-referência manual do córpus seguiu instruções para a anotação de informações de co-referência e de referências dêiticas, designadamente, elaboradas para o discurso escrito do português. A metodologia de anotação é baseada em estudos realizados nos projetos ANACORT[31], ProCaCoSA e PLN-Br[32] e conta com o uso do analisador sintático do Português PALAVRAS (Bick, 2000) e da ferramenta de anotação MMAX (*Multi-Modal Annotation in XML*) (Müller & Strube, 2001).

A anotação seguiu várias etapas: i) seleção das unidades de interesse, denominadas markables, ii) identificação de suas configurações morfossintáticas, indicação das relações entre os diversos markables, iii) classificação dos mesmos e classificação dos relacionamentos anafóricos co-referenciais e associativos.

A própria ferramenta MMAX permite codificar as marcações indicadas pelos anotadores como elementos markables, associando-os a vários atributos, conforme mostra a Tabela 1. Nessa tabela também indicamos a forma da anotação realizada, se totalmente manual (com apoio da MMAX) ou semi-automática (pelo PALAVRAS com revisão manual de sua saída). O corpus Summ-it foi anotado com informações de co-referência por uma equipe de doze anotadores, sendo que cada texto foi anotado por dois anotadores[33]. De uma forma geral, o procedimento de anotação seguiu os seguintes passos:

[29] disponível em http://nilc.icmc.usp.br:8180/portal/

[30] O projeto PLN-BR é subdividido em 7 subprojetos vinculando pesquisadores da USP, campus de São Carlos; UFSCAR; UNESP, campus de Araraquara; PUCRS; PUCRJ; UNISINOS e Universidade Presbiteriana Mackenzie.

[31] http://www.inf.unisinos.br/~renata/laboratorio/anacort_index.htm

[32] http://www.inf.unisinos.br/~renata/laboratorio/plnbr_index.htm

[33] Anotação feita em conjunto e com concordância entre os analistas humanos.

Tabela 1. Atributos dos markables

Atributos	Descrição	Forma de anotação
np_form	tipos de sintagmas nominais (Poesio, 2004)	semi-automática
pro_form	tipos de pronomes (Poesio, 2004)	semi-automática
member	indica as cadeias de co-referência (MMAX)	manual
pointer	indica uma referência associativa (MMAX)	manual
status	relações possíveis entre as entidades do discurso	manual
is_bridging	quando *status=assciative, is_bridging* indica o tipo de relação associativa	manual
is_anaphoric	quando *status=old, is_anaphoric* especifica o tipo de relação entre a entidade e o antecedente	manual
comment	usado para inserir comentários de anotação	manual

Seleção das unidades de interesse - *markables*: São os sintagmas nominais (SNs) que têm como núcleo um nome comum (*os pesquisadores*), um nome próprio (*o Museu Nacional*) ou um pronome (*Eles*). Esta etapa foi realizada de forma semi-automática. Primeiramente, os SNs foram extraídos automaticamente, com base nas informações do PALAVRAS. Após, os *markables* foram revisados manualmente utilizando a MMAX, seguindo as instruções detalhadas nos guidelines.

Identificação das configurações morfossintáticas dos *markable*: As configurações morfossintáticas são descritas pelos atributos *np form* e *pro form* , para distinguir os SNs com núcleo nome dos pronomes, respectivamente. As possíveis configurações para os sintagmas nominais com núcleo nome são:

- SNs com núcleo substantivo - **def-np**: com artigo definido (*os pesquisadores*); **indef-np**: com artigo indefinido (*um filhote*); **dem-np**: determinante demonstrativo (*essa medida*); **poss-np**: determinante pronome possessivo (*nossa pesquisa*); **int-np**:determinante interrogativo (*que horas*); **num-np**: determinante numeral (*95 empresas*); **quant-np**: com quantificadores (*v´arias respostas*); **coord-np**: coordenados (*vinho e queijo*); **bare-np**: sem determinante (*viagens*); e SNs com núcleo nome próprio **def-pn**: com artigo definido (*o Brasil*); **pn**: sem determinante (*Brasil*).

Para os pronomes temos como configurações:

- **indef-pro**: pronome indefinido (*alguém*); **dem-pro**: pronome demonstrativo (*isso*); **pes-pro**: pronome pessoal (*Eles*); **poss-pro**: pronome possessivo (*meu*); **int-pro**: pronome interrogativo (*quando*); **num-ana**: numeral ou cardinal (*Eu quero um*).

Indicação das relações entre os *markables*: Podemos anotar as relações entre os *markable* de duas formas: i) um *markable* pode indicar a retomada de outro *markable* (antecedente), quando ambos se referem à mesma entidade. Nesse caso, são co-referenciais (*o gambá - o animal*), e ligados pela relação *member* da MMAX; ii) um *markable* pode ativar um novo referente no texto cuja interpretação é dependente de um *markable* anterior, mas não se referem à mesma entidade (*macieiras - a maçã*). Quando um *markable* apresentar essa relação, o anotador deve indicar qual o *markable* que serve de âncora, pelo atributo *pointer* da MMAX.

Classificação dos markables: Nesta etapa é realizada a classificação dos SNs quanto ao seu tipo de referenciação (indicado pelo atributo *status* da MMAX). As opções são:

- ***new***: novo referente no discurso que não apresenta parte de seu sentido ancorado em uma expressão anterior (*o nordeste brasileiro*).

- ***old***: a expressão retoma um referente já introduzido por uma expressão anterior (*o gambá – o animal*).

- ***associative***: introduz um novo referente no discurso, mas cujo significado está ancorado em uma expressão anterior (*macieiras - a maçã*).

- ***deictic***: a informação requerida para interpretação da expressão não é encontrada no texto, mas na situação comunicativa (*a semana passada*).

Classificação dos relacionamentos anafóricos co-referenciais: Neste caso, temos uma subclassificação de *markables* com *status=old* e atributo *is anaphoric* em:

- ***direct***: a expressão tem um antecedente que apresenta nome núcleo idêntico (*a carta - uma carta*).

- ***indirect***: a expressão tem um antecedente que apresenta núcleo diferente (*a carta - o documento*).

- ***encapsulation***: a expressão retoma um trecho de texto maior que um sintagma, por exemplo (*a operação* retoma a sentença *O Banco Central interveio ontem para segurar a cotação do dólar*). Aqui utilizamos o atributo *comment* .

Classificação dos relacionamentos anafóricos associativos: De forma análoga à anterior, os *markables* em foco aqui são aqueles com atributo *is bridging*, que permitem a subclassificação dos *markables* do tipo *associative* nas relações seguintes (segundo as diretrizes adotadas no projeto VENEX[34]):

- ***element-of*** : a expressão anafórica é um elemento de um grupo previamente introduzido (*algumas áreas - a área*). Quando o elemento ocorre antes do conjunto, deve-se usar a relação inversa: element-of-inv (*o único dente, um molar inferior - os molares*).

- ***subset-of*** : a expressão anafórica refere-se a um subconjunto de uma entidade introduzida anteriormente no texto (*os bichos - os machos*).

- ***part-of*** : a expressão invoca parte de uma entidade já mencionada (*macieiras - a maçã*). Quando a parte ocorre antes do todo, deve-se usar a relação inversa: part-of-inv (*São Paulo - o país*).

- ***entity-attribute***: a expressão refere-se a um atributo de uma entidade previamente mencionada (*uma pesquisa com 240 casais - os resultados*).

- ***possessor-thing***: o antecedente possui a entidade evocada pela expressão associativa (*a superativação do gene - os seus efeitos colaterais*).

[34] cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf

- ***other-brigding***: outros tipos de relação não definidos pelos anteriores (*o rio - a correnteza*).

Cabe salientar que o processo de anotação de co-referência do corpus Summ-it foi dividido em duas etapas. Primeiramente, cada um dos dois anotadores realizou uma anotação inicial dos textos. Depois, cada par de anotações do texto em foco foi comparado, para se obter um consenso e, se necessário, revisar toda a anotação. Esta estratégia visou minimizar os problemas de anotação e, assim, a dificuldade da própria tarefa de anotação de co-referência.

Resultados da anotação: Os resultados da anotação de co-referência apresentados aqui foram extraídos (Collovini et al., 2007) e estão baseados no cálculo da média entre a anotação de dois anotadores para cada texto. Importante observar que esta anotação não é resultado deste projeto de mestrado, pois foi realizada pela equipe da UNISINOS, no Projeto ProCaCoSA.

Para facilitar a anotação dos 50 textos que constituem o corpus Summ-it, o mesmo foi dividido em quatro partes. A etapa de anotação de co-referência foi concluída para as quatro partes, faltando ainda o consenso final. Cabe salientar que os anotadores indicaram as relações entre os markables (atributos pointer e member) para todas as configurações dos SNs. Na anotação das cadeias de co-referência (atributo member), os anotadores identificaram um total de 526 CCRs no corpus Summ-it, tendo a mais extensa 19 membros.

Apresentamos aqui, primeiramente, a distribuição das configurações morfossintáticas dos SNs do corpus Summ-it (Tabela 2). Podemos observar que de 5050 markables, a maior parte corresponde aos SNs com nome núcleo (95,19%), pronomes sendo somente 4,81%. Devido a isso, concentramos nossa atenção somente nos SNs, mais particularmente nas descrições definidas (*np form=def-np* e *np form=def-pn*). A etapa de anotação dos markables em relação à sua anaforicidade (atributos *status, is anaphoric* e *is bridging*) levou ao seguinte resultado (Tabela 3): das 2305 descrições definidas classificadas, 1413 são da classe new (61,30%), confirmando o elevado número de informações novas nos textos. A classe *associative* representa 7,42% do total das classificações, confirmando o baixo número de casos (171) e as referências dêiticas totalizam somente 18 ocorrências

(0,78%). A classificação das descrições definidas *old* representa cerca de 30% do total de casos classificados e foi distribuída na classe *direct* que engloba 379 casos (16,44%), *indirect* com 264 casos (11,45%) e na *encapsultation* apenas 60 casos (2,60%).

Tabela 2. Resultados da identificação das configurações morfossintáticas do Summ-it

np_form	**# (%)**	***pro_form***	**# (%)**
def-np	2068 (40,95%)	indef-pro	23 (0,46%)
def_pn	386 (7,64%)	dem-pro	35 (0,69%)
indef-np	383 (7,58%)	pes-pro	152 (3,01%)
dem-np	90 (1,78%)	poss-pro	0 (0%)
poss-np	73 (1,45%)	int-pro	6 (0,12%)
int-np	2 (0,04%)	num-ana	27 (0,53%)
num-np	155 (3,07%)	**Total *pro_form***	**243 (4,81)**
quant-np	110 (2,18%)		
coord-np	98 (1,94%)		
bare-np	1134 (22,46%)		
pn	308 (6,10%)		
Total *np-form*	**4807 (95,19%)**		
TOTAL DE *MARKABLES*		**5050 (100%)**	

Tabela 3. Resultados da média da anotação de co-referência do Summ-it

Classificações		**Média (%)**
status=new		1413 (61,30%)
status=associative		171 (7,42%)
status=deitic		18 (0,78%)
status=old	*is_anaphoric=direct*	379 (16,44%)
	is_anaphoric=indirect	264 (11,45%)
	is_anaphoric=encapsulation	60 (2,60%)
Total de descrições definidas classificadas		**2305 (100%)**

Como podemos observar, o fenômeno referencial é bastante amplo e possui grande diversidade de formas referenciais, tanto nominais quanto pronominais. Neste trabalho de pesquisa, porém, restringimo-nos às formas nominais, entre as quais nos atemos às descrições definidas. Estas cobrem uma parcela considerável dos casos identificados no córpus – 2305 markables entre os 5050 identificados (45,6% do total). Deste total, 703 casos são anafóricos.

Em nossa análise, apresentamos considerações importantes principalmente no tocante aos casos de descrições definidas anafóricas do tipo indireto (264 casos).

6. Anotação manual de córpus com estruturas retóricas (RST)

A utilização de estruturas retóricas neste projeto tem por escopo não apenas a geração de sumários a partir das mesmas, como também possibilitar o estudo da estruturação RST e as possíveis relações entre o fenômeno referencial e a construção retórico-discursiva. Para tanto, foi realizada a análise RST de 12 textos do córpus Summ-it, conforme detalhamos a seguir.

6.1 Metodologia

A anotação de textos segundo a teoria RST obedeceu a um protocolo que delinearemos nesta seção. Trata-se de uma tarefa que, apesar de orientada por um conjunto pré-estabelecido de passos, insere ainda muito da subjetividade do anotador, razão pela qual a descrição detalhada das motivações das opções de anotação torna-se um repositório importante de informações a serem utilizadas em trabalhos futuros de anotação, ou mesmo na revisão do córpus anotado. A criação de repositórios de comentários referentes à anotação é de tal relevância no incremento dos trabalhos correlacionados que contamos, hoje, com ferramentas de suporte a esse tipo de análise, como a RSTTool Kit, versão melhorada da RhetDB.

A primeira etapa no processo de anotação é a segmentação dos textos, que pode ser sentencial (utilizando sinais de pontuação como delimitadores das sentenças) ou oracional. Neste último caso, o que determina que um segmento textual seja considerado uma oração depende estritamente de um protocolo determinado previamente. No caso da tarefa que apresentamos neste trabalho, foram consideradas as instruções de anotação sugeridas por Carlson e Marcu (2001), feitas para textos em língua inglesa e adaptadas para os textos em língua portuguesa.

O processo de segmentação foi feito por dois anotadores humanos, especialistas (lingüistas familiarizados com a RST), que, então, estruturaram macro e micro-estruturalmente os textos, valendo-se do auxílio de uma ferramenta de suporte à anotação, a RSTTool (O´Donnell, 2000). Note-se que a referida ferramenta não automatiza a análise em nenhum

aspecto, servindo apenas como um ambiente amigável para a anotação, e que fornece recursos gráficos úteis para a visualização da estrutura arbórea decorrente da anotação.

Para estruturar o texto, os anotadores identificam possíveis relações RST às suas unidades. As relações da estruturação do texto são funcionais, ou seja, o que importa é a categoria do efeito que elas produzem. Elas podem ser descritas em termos das finalidades do produtor textual, das suas suposições sobre o leitor, e de determinados padrões proposicionais em relação ao conteúdo do texto. "As relações da estruturação do texto refletem as opções do produtor de organização e apresentação; é nesse sentido que a RST é 'retórica'" (Mann; Matthiessen; Thompson, 1992, p. 45).

Para atribuir relações às unidades, os anotadores se valeram das definições das relações RST e observam se a definição plausivelmente se aplica às unidades em questão. Um exemplo de definição de relação é o seguinte (Mann; Matthiessen; Thompson, 1992)[35]:

Nome da relação: EVIDENCE

Condições no núcleo (N): o leitor pode não acreditar no núcleo em um grau de satisfação para o produtor textual.

Condições no satélite (S): o leitor acredita no satélite ou o acha crível.

Condições na combinação núcleo-satélite (N + S): a compreensão do leitor do satélite aumenta sua crença no núcleo.

Efeito: a crença do leitor no núcleo é aumentada.

***Locus* do efeito**: núcleo.

Cada campo de uma definição de relação especifica julgamentos particulares que os anotadores do texto devem fazer na construção da estrutura RST. Os anotadores têm acesso ao texto, têm conhecimento do contexto no qual ele foi escrito e compartilham as convenções culturais do produtor textual e dos leitores pretendidos, mas não têm acesso

[35] Tradução nossa.

direto nem ao produtor textual nem a outros leitores. Por isso, seus julgamentos devem ser de plausibilidade.

Na Figura 19, podemos observar a análise RST de um trecho de texto diagramada na RSTTool, que ilustra a relação EVIDENCE, definida acima:

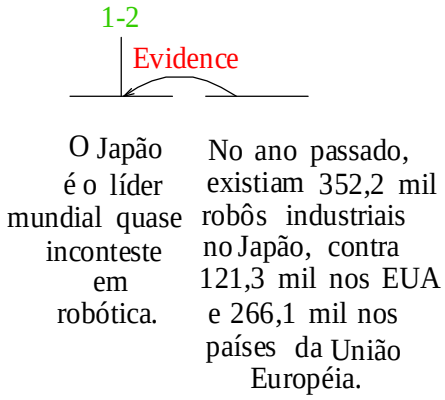


Figura 19. trecho de texto do córpus Summ-it analisado na RSTTool

Na RSTTool, cada segmento de uma relação é marcado por uma linha horizontal. O segmento nuclear é marcado por uma linha vertical, e a relação é marcada por uma seta, que aponta sempre para o núcleo, exceto no caso de relações multinucleares.

Conforme vimos anteriormente, a RST permite a adoção de conjuntos determinados de relações retóricas. Na anotação que procedemos, foi utilizado o conjunto proposto por Pardo (2005b). O conjunto de Pardo (2005b) foi adotado por ser o utilizado no analisador discursivo automático DiZer (Pardo, 2005b), para o qual foi feita uma análise de córpus (textos científicos do domínio da Ciência da Computação) com o objetivo de identificação das relações retóricas presentes nos textos em português, bem como seus respectivos indicadores, ou seja, marcadores discursivos e expressões indicativas.

6.1.1 Segmentação

Para identificar as relações em um texto, Mann e Thompson (1987) sugerem que o analista deve primeiramente segmentá-lo em unidades textuais ou, segundo Carlson e Marcu (2001), unidades discursivas elementares, aqui representadas pela sua sigla em inglês, EDUs (*Elementary Discourse Units*). Uma EDU é considerada um bloco mínimo de construção de uma árvore discursiva (sempre ocorrendo como uma folha da árvore), correspondendo a uma proposição elementar, no nível discursivo. O tamanho da unidade de discurso é arbitrário para a RST, podendo abranger desde itens lexicais típicos até parágrafos inteiros ou unidades ainda maiores. Porém, as unidades devem ter integridade funcional independente. Carlson e Marcu (2001) sugerem que se considere a oração como a unidade elementar do discurso, usando indícios lexicais e sintáticos para ajudar na determinação de fronteiras. Após delimitar EDUs no texto, o passo seguinte consiste em estabelecer as relações entre elas.

Durante o processo de segmentação dos textos do córpus Summ-it, porém, houve pontos passíveis de dúvidas, para os quais estabelecemos algumas diretrizes:

a) consideramos Orações Reduzidas (adverbiais e relativas não-restritivas) como EDUs.

Ex: O estudo, feito por pesquisadores do Imperial College, em Londres, mostra que (...).

b) não consideramos Orações Restritivas como EDUs.

Ex: (...) os pesquisadores analisaram o fígado de mulheres que haviam sofrido um transplante de medula óssea (...).

c) segmentos, mesmo que não oracionais, que estão entre parênteses, travessões ou outros delimitadores gráficos são considerados EDUs e relacionados através da relação PARENTHETICAL.

Ex: A projeção dos cientistas para o ano 2025 é que 3,3 bilhões de pessoas não tenham mais água para irrigação – a atividade humana que mais consome o líquido.

Ex: A tecnologia difere do biodiesel utilizado em outras partes do mundo, que usa o metanol – um derivado do petróleo.

Em princípio, as EDUs devem ser só oracionais. A relação PARENTHETICAL possibilita que segmentos não oracionais sejam considerados EDUs, desde que sejam separados por delimitadores gráficos "fortes". Para Marcu (2001), as vírgulas seriam delimitadores "fracos" (não fortes o suficiente para fazer de uma não-oração uma EDU). Porém, nem sempre os delimitadores gráficos indicam a relação PARENTHETICAL. É o caso de segmentos como: "Pessoas nascidas na China têm mais facilidade de se lembrar de um objeto quando o vêem pela segunda vez com o mesmo fundo que aparecia na primeira olhada – o que já não acontece com os americanos." Nesse caso, o segmento após o travessão estabelece uma relação CONTRAST com os segmentos anteriores.

Por outro lado, alguns segmentos que não estão separados por delimitadores gráficos "fortes" podem estabelecer uma relação PARENTHETICAL. É o caso de segmentos como "O estudo, feito por pesquisadores do Imperial College, em Londres, mostra que (...)", em que o segmento entre vírgulas se encaixa como satélite na descrição da relação PARENTHETICAL: "S apresenta informação extra relacionada a N, complementado N; S não pertence ao fluxo principal do texto".

d) segmentos, mesmo que não oracionais, que indicam a autoria de discursos diretos ou indiretos, são considerados EDUs e relacionados através da relação ATTRIBUTION. Isso implica assumir que opiniões de autoria, acompanhadas da indicação do autor, quer direta (em transcrições literais), quer indiretamente (como no uso de "segundo fulano..."), são particionadas em duas EDUs. Carlson & Marcu (2001) fala em "reported speech" e "reporting speech".

Ex: "É uma descoberta e tanto", disse o psicólogo César Ades.

Ex: Segundo Kellner, apesar de o animal ser um baixinho, (...).

Ex: O ministro da Agricultura, Roberto Rodrigues, afirmou que o uso da soja transgênica "é uma boa idéia".

Além dessas diretrizes específicas adotadas pelos anotadores para a segmentação oracional, também foi decidido consensualmente como lidar com outros aspectos dos textos do córpus, como títulos, por exemplo. Como o córpus é composto de textos de divulgação científica publicados em contexto midiático, a sua grande maioria, além dos títulos, apresenta subtítulos, que aparecem geralmente entre porções de texto. Decidimos desconsiderar tanto os títulos quanto os subtítulos dos textos, suprimindo-os ao colocar os textos na RSTTool. Porém, registramos cada caso de supressão, para que haja um controle da maneira como estamos tratando os textos do córpus.

6.1.2 Estruturação

Em cada texto, após a segmentação, devem ser atribuídas relações às EDUs consideradas nucleares e satélites. As relações entre as EDUs formam segmentos maiores, entre os quais são atribuídas relações, formando segmentos maiores ainda, e assim sucessivamente, até a constituição da estrutura hierárquica completa do texto.

Essa estrutura hierárquica pode ser observada na Figura 20.

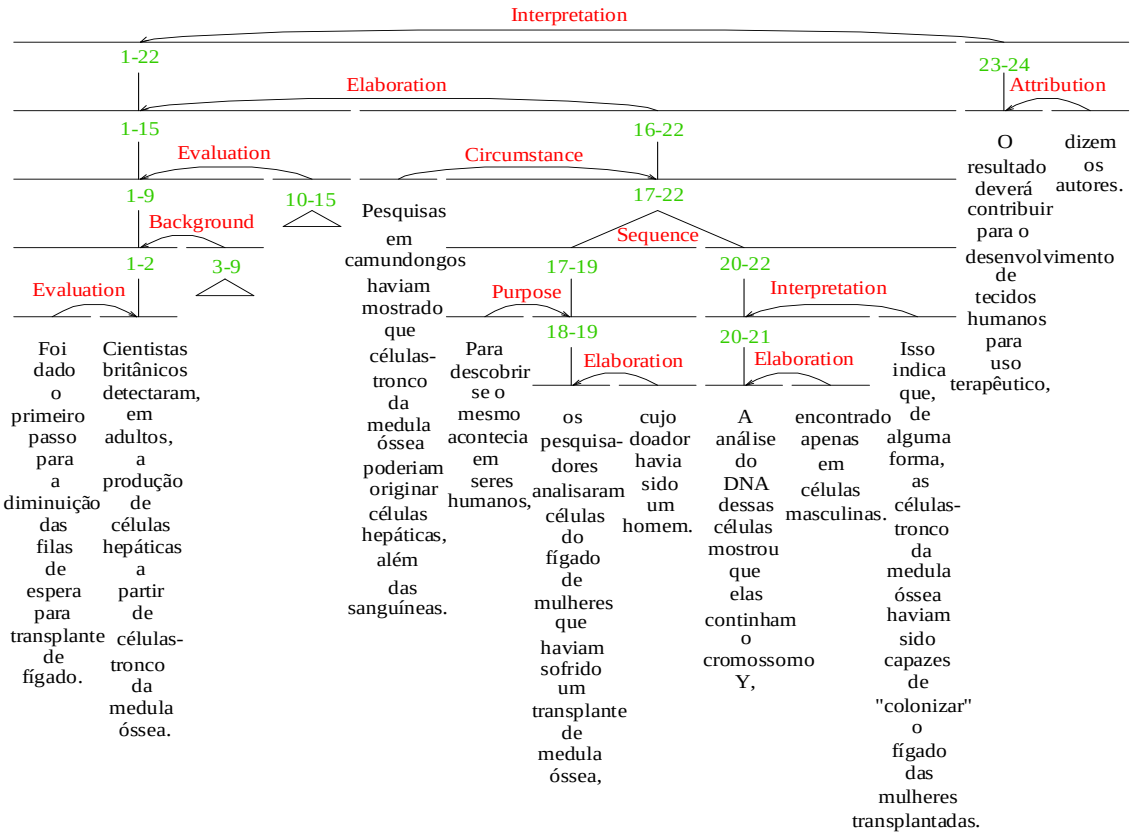


Figura 20. Análise do texto CIENCIA_2000_17109, diagramada na RSTTool

Na estrutura hierárquica apresentada como exemplo, é possível observar que as relações são atribuídas recursivamente aos diversos níveis hierárquicos do texto. A relação ELABORATION, por exemplo, estabelece-se no nível elementar (das EDUs), como entre os segmentos 18 e 19 ou entre os segmentos 20 e 21, e também em um nível mais superior do texto, macroestrutural, como entre os segmentos 1-15 e 16-22.

Para atribuir relações aos segmentos textuais, conforme dito anteriormente, os anotadores consultam as definições das relações e observam se a relação plausivelmente se aplica aos segmentos em questão. Porém, alguns tipos de segmentos dos textos analisados geraram dúvidas sobre que relação aplicar. A partir da discussão sobre essas dúvidas, os anotadores definiram consensualmente algumas diretrizes sobre que relação aplicar em cada caso.

A seguir, são comentados os resultados encontrados na análise de 12 textos do córpus Summ-it, bem como as diretrizes consensuais adotadas pelos anotadores para os casos duvidosos de aplicação das relações.

6.2 Resultados

A incidência das relações encontradas numa amostra de 12 textos do córpus Summ-it pode ser observada na Tabela 4:

Tabela 4. Incidência das relações RST encontradas na análise de 12 textos do córpus Summ-it

Relações	# (%)	Relações	# (%)
Elaboration	93 (23,25%)	Sequence	7 (1,75%)
Attribution	50 (12,50%)	Evidence	5 (1,25%)
Parenthetical	42 (10,50%)	Non-volitional-result	5 (1,25%)
Same-unit	34 (8,50%)	Conclusion	4 (1,00%)
Interpretation	26 (6,50%)	Joint	3 (0,75%)
Evaluation	20 (5,00%)	Antithesis	2 (0,50%)
Purpose	20 (5,00%)	Explanation	2 (0,50%)
Background	19 (4,75%)	Means	2 (0,50%)
List	17 (4,25%)	Non-volitional-cause	2 (0,50%)
Circumstance	15 (3,75%)	Otherwise	2 (0,50%)
Contrast	10 (2,50%)	Comparison	1 (0,25%)
Condition	9 (2,25%)	Justify	1 (0,25%)
Concession	8 (2,00%)	Solutionhood	1 (0,25%)
Total de incidências das relações de 12 textos:			**400** (100%)

Algumas dessas relações foram encontradas principalmente no nível elementar (entre as EDUs) dos textos. Além disso, essas relações, em sua maioria, apresentaram marcadores textuais que as identificavam. A Tabela 5 apresenta a incidência de relações que apareceram no nível elementar dos textos, e a porcentagem, em cada relação, de relações que foram identificadas por marcadores textuais.

Tabela 5. Marcadores indicativos de relações

RELAÇÃO (número de vezes em que ocorreu no total)	Número de vezes em que ocorreu no nível elementar dos textos	Número de vezes em que apresentou marcadores textuais	Marcadores textuais observados (no início ou no meio do satélite)[36]
Elaboration (93)	50	36	pronomes relativos (que-13, onde-4, cujo-2, como-1, em que-1); verbos no particípio-9 (assinadas, encontrado, aprimorados, todas causadas, escolhido, realizado e promovido, desenvolvido, ocultas, encerrado); verbos no gerúndio-4 (forçando, fartando-se, prestando atenção, gesticulando); advérbios-2 (anualmente, literalmente)
Attribution (50)	50	48	verbos (disse, diz, dizem-20, afirmou, afirma-14, conta, contou-2, sugerem-2, explicou-1, argumentaram-1, descobriu-1, defendeu-1, resume-1); conjunções (segundo-4); preposições (para-1)
Patenthetical (42)	42	42	parênteses-27; travessões-10; colchetes-4; verbo no particípio (feito-1)
Same-Unit* (34)	34	31	parênteses-16; travessões-4; pronomes relativos (que-3); conjunções (segundo-2, quando-1); verbos (disse-1, forçando-1, desenvolvido-1, feito-1, aprimorados-1)
Purpose (20)	20	20	preposições (para-18); conjunções (com o objetivo de-1, na tentativa de-1)
Circumstance (15)	12	11	conjunções (quando-4; sem-2; ao-2; enquanto-1; assim que-1); pronome relativo (onde-1)

[36] *Nas relações multinucleares, os marcadores apareceram em um dos núcleos. Na relação SAME-UNIT, os marcadores se referem ao satélite que se interpôs entre os núcleos unidos.

List* (17)	10	10	conjunções (e-10)
Concession (8)	8	8	conjunções (mas-6, apesar de-2)
Condition (9)	7	7	conjunções (se-5); verbos no particípio (não satisfeita-1, dopada-1)
Evaluation (20)	7	5	adjetivos (menos precisos-1; minúsculas-1, incompleto-1); verbos (não pode-1, parece até-1)
Background (19)	4	1	verbo no particípio (batizado-1)
Non-Volitional Result (5)	4	4	verbos no gerúndio (aumentando-1, matando-a-1, destruindo-1); preposição (até-1)
Contrast* (10)	4	2	conjunções (enquanto-2)
Otherwise (3)	3	3	conjunções (em vez de-3)
Interpretation (26)	3	2	expressões (é como se-1, isso indica que-1)
Antithesis (2)	2	2	conjunções (mas-2)
Means (2)	2	2	verbos no gerúndio (olhando-1, usando-1)
Non-Volitional Cause (2)	2	2	preposição (por-1); expressão (se deve a-1)
Sequence* (7)	1	1	conjunção (e-1)
Conclusion (4)	1	0	
Justify (1)	1	0	
Explanation (2)	1	0	
Total de relações: 400	**Total de relações que ocorreram no nível elementar dos textos: 268**	**Total de relações do nível elementar que apresentaram marcadores textuais: 237**	

Conforme mostra a Tabela 5, 67% das relações dos 12 primeiros textos do corpus Summ-it apareceram no nível elementar dos textos; e, destas, 88,4% apresentaram marcadores textuais. É possível observar que há relações que apresentam todas as (ou a maioria das) suas ocorrências no nível elementar dos textos. É o caso das relações ATTRIBUTION, PARENTHETICAL, SAME-UNIT, PURPOSE, CIRCUMSTANCE, LIST,

CONCESSION, CONDITION, NON-VOLITIONAL RESULT, OTHERWISE, ANTITHESIS, MEANS, NON-VOLITIONAL CAUSE, JUSTIFY e EXPLANATION.

Por outro lado, há relações que dificilmente aparecem no nível elementar dos textos. É o caso das relações EVALUATION, BACKGROUND, CONTRAST, INTERPRETATION, SEQUENCE e CONCLUSION. Essas relações foram encontradas quase que exclusivamente em níveis mais superiores, macroestruturais, dos textos. Além disso, essas relações quase não apresentaram marcadores textuais que as identificassem. Um caso ímpar é o da relação ELABORATION, que, por sua versatilidade, aparece tanto no nível elementar quanto no nível macroestrutural dos textos.

Algumas relações, por sua vez, foram encontradas quase que exclusivamente em níveis mais superiores, macroestruturais, dos textos. Além disso, essas relações não apresentaram marcadores textuais que as identificassem.

A incidência de determinadas relações nos níveis macroestruturais dos textos está ligada à superestrutura do gênero de texto em questão – de divulgação científica publicado em contexto midiático. Nesse gênero textual, diferentemente do que acontece no texto científico dirigido aos pares (publicado em periódicos científicos de uma dada comunidade científica), é razoável afirmarmos que não se espera que o leitor esteja interessado de antemão na pesquisa que será veiculada. Além disso, nos meios de comunicação de alta circulação – como jornais e revistas –, a matéria é um produto a ser comercializado. Portanto, o texto de divulgação científica publicado em contexto midiático deverá, antes de tudo, captar o leitor.

É por isso que se podem observar, nos textos do córpus Summ-it, especialmente no início de cada texto, trechos que parecem desempenhar essa função de captação do leitor. É o caso dos seguintes trechos iniciais de textos:

- texto CIENCIA_2000_17108: “Um ser que invade corpos e domina a mente alheia, forçando suas vítimas a fazer o que ele ordena, não é mero personagem de ficção. Para uma aranha da Costa Rica, essa criatura existe”;

- CIENCIA_2000_17109: “Foi dado o primeiro passo para a diminuição das filas de espera para transplante de fígado”;

- texto CIENCIA_2000_17112: “O mundo está mais seco do que se imaginava”;

- texto CIENCIA_2004_26415: “Para um desavisado parece até obsessão freudiana, mas Hendrik Poynar está pedindo a todos os seus conhecidos a maior quantidade de fezes possível – quanto mais velhas, melhores”;

- texto CIENCIA_2005_28747: “Chineses e americanos enxergam o mundo de jeitos distintos – literalmente, a julgar por uma pesquisa publicada hoje”.

Essa função de captação do leitor é pragmática. Portanto, os trechos de texto exemplificados seriam satélites de alguma relação intencional (ou apresentativa), cujo efeito fosse aumentar alguma inclinação no leitor para a leitura do núcleo (e cujo *locus* do efeito estivesse apenas no núcleo). Pensamos primeiramente na relação PREPARATION[37], que, como se pode observar pela sua descrição e pelo seu uso em Mann (2006), seria aplicável a esses trechos. Porém, essa relação não figura entre as relações do conjunto adotado neste trabalho (Pardo, 2005b). Como entre as relações do conjunto de Pardo (2005b) não há nenhuma relação intencional cuja descrição pareça se aplicar aos trechos em questão, optamos pelo uso das relações semânticas EVALUATION[38] e INTERPRETATION[39]. Essas relações, embora semânticas, apresentam, na sua descrição, elementos que evidenciam a posição do produtor do texto, no satélite, frente ao que é dito no núcleo.

Outros elementos da superestrutura do gênero texto de divulgação científica publicado em contexto midiático, que podem ser observados na análise dos textos do córpus Summ-it, são: (i) a menção à pesquisa divulgada; (ii) a apresentação dos procedimentos metodológicos utilizados na pesquisa divulgada; (iii) a avaliação e a interpretação dos

[37] Efeito: o leitor se sente mais preparado, interessado ou orientado para ler o núcleo.
[38] Efeito: o leitor reconhece que o satélite avalia a situação apresentada no núcleo e reconhece o valor que lhe é atribuído.
[39] Efeito: o leitor reconhece que o satélite relaciona o núcleo com um quadro de idéias não envolvidas no conhecimento apresentado no próprio núcleo.

pesquisadores ou de outros membros da comunidade científica sobre a repercussão da pesquisa; (iv) a menção a trabalhos futuros ou relacionados à pesquisa divulgada; (v) a menção a outros periódicos em que a pesquisa foi divulgada. Para cada um desses elementos superestruturais, convencionamos atribuir determinadas relações.

A menção à pesquisa divulgada é sempre o segmento nuclear mais saliente na estrutura hierárquica dos textos, já que a finalidade do gênero textual em questão é divulgar alguma pesquisa. Todas as relações macroestruturais de cada texto (subárvores internas) estão ligadas como satélites a esse segmento nuclear, conforme podemos observar na Figura 21.

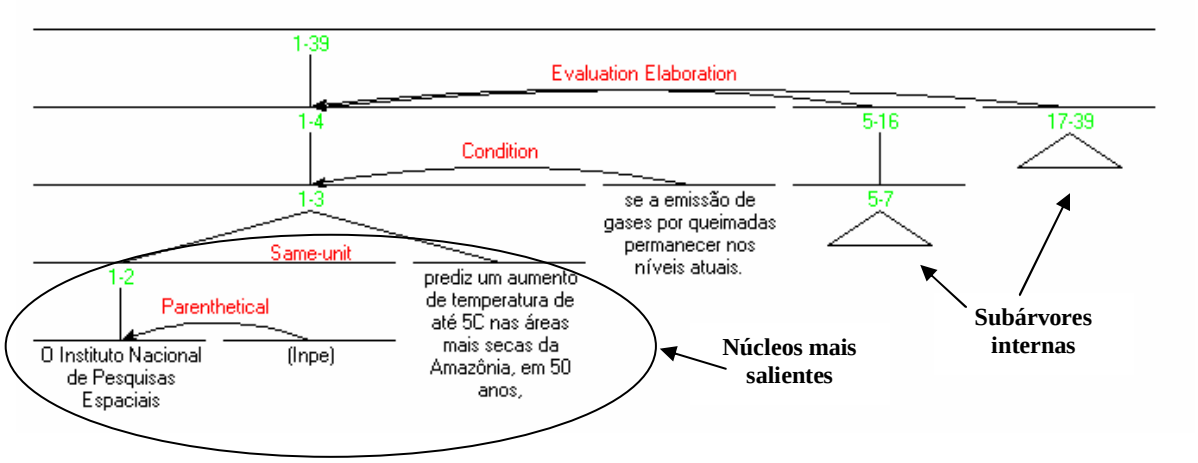


Figura 21. Estrutura RST do texto CIENCIA_2000_17082

A apresentação dos procedimentos metodológicos utilizados na pesquisa divulgada geralmente constitui um segmento macroestrutural diretamente ligado ao segmento nuclear mais alto na estrutura hierárquica do texto. A relação mais freqüente para esse tipo de segmento é ELABORATION.

A avaliação e a interpretação dos pesquisadores ou de outros membros da comunidade científica sobre a repercussão da pesquisa geralmente constitui um segmento macroestrutural diretamente ligado ao segmento nuclear mais alto na estrutura hierárquica do texto (vide figura 21). As relações mais freqüentes para esse tipo de segmento são EVALUATION e INTERPRETATION.

A menção a trabalhos futuros ou relacionados à pesquisa divulgada, bem como a menção a outros periódicos em que a pesquisa foi divulgada, constituem segmentos macroestruturais

que costumam aparecer no final de cada texto, ligados ao restante do texto. A relação mais freqüente para esse tipo de segmento é ELABORATION, podendo ser também MOTIVATION. É o que temos nas Figura 22 e Figura 23 a seguir.

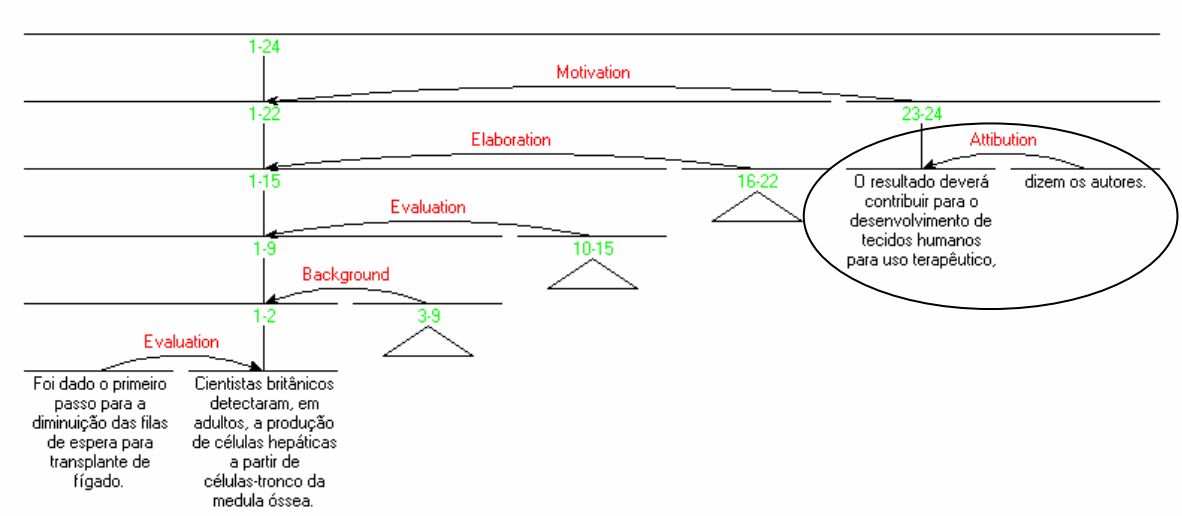


Figura 22. Estrutura RST do texto CIENCIA_2000_17109

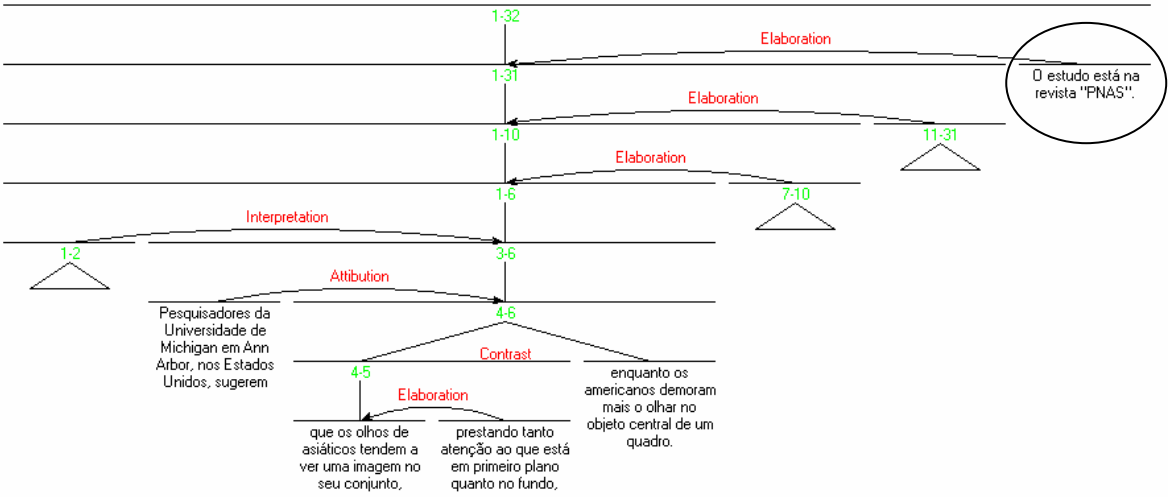


Figura 23. Estrutura RST do texto CIENCIA_2005_28747

7. Sistemas automáticos utilizados

Nesta seção, descrevemos os sistemas computacionais que são utilizados neste projeto de pesquisa e já apontados no preâmbulo a esta parte do trabalho (vide figura 17).

7.1 DiZer

O DiZer (Pardo, 2005b) é, hoje, o único *parser* (analisador) discursivo automático para o português. Este sistema faz uso de padrões de análise lingüística, manualmente extraídos de um córpus de textos científicos na área de Ciência da Computação com a finalidade de identificar e construir árvores RST. São aproximadamente 750 padrões de análise que dão conta de, a partir de palavras e expressões indicativas, identificar as relações discursivas presentes no texto. Este processo, segundo Marcu (1999), é tradicionalmente considerado como a ferramenta lingüística mais importante para a identificação de estruturas retóricas em um texto.

Sua arquitetura está descrita na Figura 24 (adaptada de Pardo, 2005b):

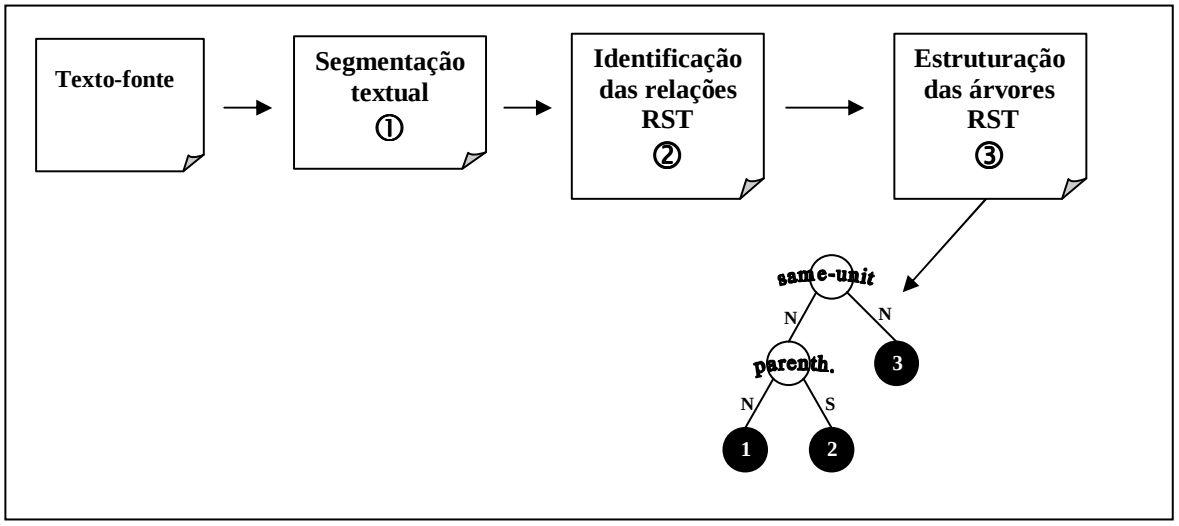


Figura 24. Arquitetura do DiZer

Como podemos observar, o sistema está ancorado no encadeamento de três processos: i) segmentação do texto-fonte; ii) identificação das relações no mesmo; e iii) estruturação das árvores, saída do sistema.

No processo de segmentação, o DiZer tenta determinar as sentenças no texto-fonte, geralmente expressas por unidades discursivas (proposições, no âmbito da RST). Para tanto, o sistema utiliza uma marcação morfossintática que etiqueta cada palavra do texto, feita por um etiquetador para o Português brasileiro (Aires, 2000), e segmenta o texto de acordo com a identificação de sinais de pontuação (pontos finais, vírgulas etc.) ou quando uma palavra ou expressão indicativa forte é identificada. Por conta da ambigüidade do ponto, que pode ser sinal de pontuação ou de abreviação, o sistema conta com uma lista de termos abreviados que usam o sinal. As palavras ou expressões indicativas são aquelas que possuem uma função identificada no discurso e não são ambíguas[40] (*mas, portanto, ou seja* etc.), indicando, assim, uma relação discursiva, e pertencem a um repositório de termos indicativos. Além desses procedimentos, o sistema ainda verifica se o segmento é oracional através da identificação de um verbo no mesmo.

No tocante à identificação (determinação) das relações retóricas entre os segmentos do texto, o DiZer procura relacionar, sempre, dois segmentos adjacentes – gerando estruturas binárias, ou seja, composta por duas proposições[41]. Inicialmente, o sistema relaciona os segmentos adjacentes; então, relaciona estes spans dentro do parágrafo a que pertencem; e, ao final, relaciona os parágrafos adjacentes. Esta ordem de processamento baseia-se na premissa de que o escritor distribui a informação transmitida no texto em níveis de organização iguais (a idéia de progressão textual, segundo a qual as idéias de organizam em parágrafos, e que estes possuem relações retóricas entre si). É o que acorre nos casos em que duas proposições expressam, respectivamente, causa e conseqüência:

[O Governo enrijeceu a base no Congresso]$_1$ [***por causa*** das ofensivas da oposição.]$_2$

Os segementos 1 e 2 pertencem a um mesmo nível de organização e podem ser relacionados como CAUSA(1_N,2_S), onde 2 (satélite) é causa de 1 (núcleo). O sistema identifica esta relação através do já referido repositório, que conta com diversos "padrões discursivos" identificados por Pardo através de análise de córpus (textos científicos da área

[40] Palavras como *e* e *se*, que podem ter funções morfossintáticas variadas, são ignoradas.
[41] O DiZer não gera estruturas com mais de duas proposições unidas por um mesmo nós estrutural, mesmo nos casos de relação multinuclear (uma relação LIST, por exemplo, composta por três ou mais EDUs).

da Ciência da Computação). Na ausência de marcadores identificáveis, o sistema considera a relação entre dois segmentos ou subárvores como uma ELABORATION.

Importante ressaltar, que em um sistema como o DiZer o nível de dependência entre o domínio dos textos de treinamento e o desempenho tende a ser alto. Por conta disso, mesmo, a opção por textos de entrada pertencentes a outro domínio não pode precindir de um novo treinamento do sistema, através de um novo estudo de córpus.

Essas características do sistema serão recuperadas a frente (seção 8), onde apresentamos uma avaliação do acoplamento do DiZer ao RheSumaRST, sistema de SA que passamos a descrever.

7.2 RheSumaRST

O RheSumaRST (Seno, 2005; Seno & Rino, 2005) é um sumarizador automático profundo baseado na RST que tem como entrada uma representação da estrutura retórica do texto-fonte (tarefa de análise discursiva previamente feita) – que, no caso desse sistema, é a árvore RST do texto-fonte – a partir da qual determina o conteúdo e a forma de seus possíveis sumários (tarefa de redução), ou seja, produz a estrutura retórica do sumário correspondente. Nesse modelo baseado na RST, a redução explora a **assimetria do relacionamento proposicional**, pela identificação de funções discursivas distintas.

A opção por este protótipo de SA deveu-se, além da modelagem baseada em RST e anotação de veias, na possibilidade de acoplamento ao DiZer. Uma parte importante do trabalho desenvolvido neste projeto de mestrado está relacionada ao estudo dos primeiros resultados dessa junção (que passamos a apresentar na seção 8).

O sistema pode ser representado por uma arquitetura em *pipeline*, conforme vemos na Figura 25:

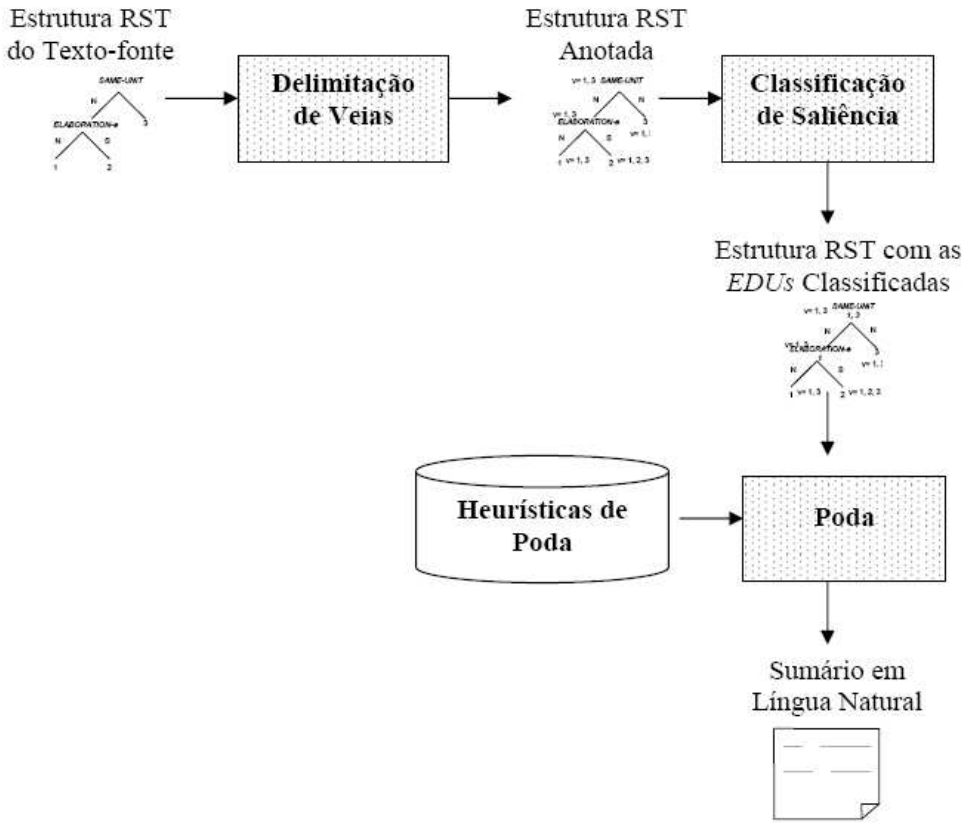


Figura 25. Arquitetura do RHeSumaRST

Observemos que, na arquitetura acima descrita, a entrada do sistema não é um texto-fonte em língua natural, mas uma representação do mesmo (a representação conceitual I que temos na arquitetura da Figura 3). A geração dessa representação conceitual é tarefa desenvolvida por um *parser* discursivo que gera, a partir do texto-fonte em língua natural, árvores RST representativas das relações retóricas entre as proposições do mesmo.

Conforme já explanado na 5.1, o texto pode ser segmentado em unidades discursivas mínimas (*EDUs*) que se relacionam, sendo identificáveis as proposições nucleares, ou seja, mais relevantes. Observemos, porém, que o RheSumaRST não seleciona a informação mais relevante para o sumário apenas agrupando proposições nucleares das árvores RST submetidas ao sistema. Antes do processo de redução, as árvores são submetidas a um processamento algorítmico que faz a delimitação do domínio de acessibilidade referencial do texto, ou seja, que liga todas as proposições (folhas da árvore RST) com outras que

possam fazer referência às mesmas. Esse processo é, como foi dito, algorítmico e baseado na Teoria das Veias.

Ao ser feito este processamento, o sistema retorna, para cada folha da árvore RST, uma informação que indica quais outras folhas são prováveis co-referentes. Assim, para uma dada folha X, o sistema retornaria dados no seguinte formato: head = X, veins = Y, Z, T, onde Y, Z e T são outras folhas às quais X pode fazer referência.

Após a delimitação das veias, então, ocorre o processo de sumarização em si, gerando uma nova estrutura RST que é a representação conceitual do sumário (representação conceitual II – vide figura 1). Este processo de redução é baseado em uma heurística de poda desenvolvida a partir de um estudo de córpus. Esta heurística visa à identificação de *EDU* menos relevantes na estrutura RST de um texto, o que permite excluir somente aquelas que não interfiram na coerência, isto é, as *EDUs* candidatas a exclusão serão somente aquelas que estão fora do domínio de acessibilidade referencial de outras *EDU* candidatas a inclusão. As heurísticas baseiam-se na verificação da acessibilidade referencial de cada *EDU* já inclusa na estrutura do sumário. Seno & Rino (2005) descrevem a heurística de poda juntamente com sua descrição funcional. Observemos a descrição[42]:

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

Na avaliação de coerência feita por Seno, o RheSumaRST foi comparado com o Modelo de Saliência de Marcu (1997), e com um *baseline* desenvolvido especificamente para o experimento, o Topline, que apenas seleciona os núcleos das árvores RST, podando-os da esquerda para a direita na estrutura. Para um córpus de 10 textos jornalísticos[43]

[42] Os satélites são apresentados em negrito. Para simplificação, supõe-se, nesses exemplos, que os satélites não pertencem às veias de outras EDUs.
[43] Extraídos do Córpus TeMário - http://www.linguateca.pt/Repositorio/TeMario/.

(aproximadamente uma página e meia cada um), os resultados da avaliação manual dos sumários gerados estão representados na Tabela 6 (Seno, 2005).

Tabela 6. Avaliação de coerência do RheSumaRST

SISTEMAS	# de CCRs	# de quebras de CCRs	Quebra de CCRs (%)
Saliência	81	12	15
Topline	89	7	**8**
RHeSumaRST	**93**	**5**	**5**

Em um segundo experimento de avaliação, com metodologia idêntica, o sistema foi testado com um córpus de 25 textos menores[44] (meia página cada, em média). O interessante neste experimento é a avaliação do sistema não apenas para textos jornalísticos, mas também para os científicos (Seno, 2005).

Tabela 7. Avaliação do RheSumaRST para textos jornalísticos e científicos

	SISTEMAS	# de CCRs	# de quebras de CCRs	Quebra de CCRs (%)
JORNALÍSTICO	*Topline*	45	8	18
	RHeSumaRST	45	5	4
CIENTÍFICO	SISTEMAS	# de CCRs	# de quebras de CCRs	Quebra de CCRs (%)
	RHeSumaRST	17	5	29
	Topline	23	5	22

Como podemos observar, apesar de, em tese, o sistema ser independente do domínio ao qual o texto pertence, o desempenho do sistema é melhor para textos jornalísticos. Esse

[44] Extraídos do Córpus Rhetalho - http://www.icmc.usp.br/~taspardo/rhetalho.html.

dado é importante, pois orientará nossas decisões com relação ao experimento 01 (E1), que passamos a descrever na próxima seção.

8. Estudo de viabilidade de sistemas e validação de teorias para a Língua Portuguesa

Os experimentos realizados no âmbito deste projeto de mestrado visaram, primeiramente, ao fornecimento de subsídios lingüísticos para a melhoria dos sistemas de Sumarização Automática, principalmente no tocante à manutenção das cadeias de co-referência, mas também com relação à compreensão de outros elementos de considerável importância no âmbito da automação do processamento de língua natural, como é o caso da questão da influência do gênero textual no processamento de textos. Os dados de análise lingüística derivados das observações feitas nos experimentos que são descritos nesta seção são, portanto, a base das efetivas contribuições deste trabalho de mestrado.

Nesse sentido, para atingir os objetivos de pesquisa propostos, nossa metodologia incluiu a utilização de córpus anotados com estruturas retóricas (RST) e cadeias de co-referência e baseou-se tanto na análise destas anotações quanto na análise dos sumários gerados automaticamente. A partir deste estudo analítico, foram diagnosticados problemas nos processos automáticos, o que permitiu a proposição e implementação de melhorias no processo de sumarização automática do modelo em estudo, o que veremos nas seções adiante.

Outro aspecto importante dos experimentos, são as contribuições na forma de ferramentas e sistemas implementados através de um importante trabalho de parceria e interação entre o lingüista e o cientista da computação. Como veremos nos resultados descritos, o intercâmbio de informações e idéias entre esses dois profissionais parece ser um canal de produção extremamente interessante e frutífero, pois alia esferas de conhecimento distintas na sua especificidade, porém complementares na consecução dos objetivos específicos do PLN.

8.1 Experimento 01 – RheSuma-2: acoplamento do DiZer ao RheSuma-RST

O RheSuma-2 (Carbonel et al., 2006) consiste basicamente na junção dos processamentos feitos pelos dois sistemas descritos anteriormente: o DiZer e o RHeSumaRST. O objetivo dos

experimentos realizados, então, foi avaliar os sumários produzidos, focando esta avaliação nas quebras de CCR. Através do estudo pontual das mesmas, pretendeu-se elencar os pontos problemáticos na modelagem computacional das CCR e, posteriormente, aproveitar estas observações para melhoras nos sistemas existentes, bem como nos protocolos de anotação de CCR.

Como vimos, o DiZer foi desenvolvido e treinado para textos científicos e o RHeSumaRST, apesar de melhores resultados com textos jornalísticos, não possui dependência de domínio. Assim, o natural seria que o córpus escolhido para este experimento fosse de textos científicos. Todavia, a fim de inserir o experimento no contexto mais amplo do Projeto ProCaCoSA, optou-se pelo texto jornalístico. Isso, então, implicou a penalização do desempenho do DiZer, conforme veremos nas considerações tecidas adiante.

Para o experimento[45], foi utilizado o córpus Rhetalho[46], que conta, originalmente, com 40 textos anotados retoricamente por dois anotadores humanos. Para o experimento em foco foram utilizados 47 textos que não haviam sido submetidos ao RHeSumaRST anteriormente. Com uma média de 200 palavras cada – o correspondente a uma manchete curta no jornal, ou seja, cerca de meia página de texto – estes textos foram coletados junto ao site do Jornal A Folha de São Paulo e pertencem a seções distintas do jornal (10 textos da seção Cotidiano; 9 textos da seção Ciência, 9 textos da seção Esporte; 10 textos da seção Informática e 9 textos da seção Mundo).

A fim de possibilitar a avaliação do desempenho do sistema no tocante às CCR, o córpus passou por um processamento em vários níveis: i) pré-processamento por um *parser* sintático para o português (PALAVRAS – Bick, 2000); ii) anotação das cadeias de co-referência, observados apenas os casos de descrições definidas remetendo a um termo antecedente; iii) geração de um conjunto de arquivos para cada texto analisados (arquivos de *words*, *markables*, *tokens*, *chunks* e de anotação). Através destas informações anotadas foi possível um mapeamento das ocorrências de cadeias de co-referência nos textos-fonte e, posteriormente, o cômputo do número de quebras nos sumários gerados. Ressalte-se que, apesar de haver várias

[45] Mais bem detalhado em Carbonel & Rino, 2006 – relatório técnico.
[46] http://www.icmc.usp.br/~taspardo/rhetalho.html

modalidades de cadeias de co-referência, o foco deste experimento são as descrições definidas, não sendo anotadas as demais (pronominais, associativas etc.).

Para a avaliação do desempenho do RheSuma-2, foram utilizados dois baselines: o Topline (Seno & Rino, 2005), que, dada uma árvore RST, seleciona apenas os núcleos da mesma e usa como indicador de saliência a mera posição das EDUs – ordena-a de cima para baixo, e da esquerda para a direita; e o Modelo de Saliência (Marcu, 1997), um algoritmo que calcula e atribui pesos às EDU da estrutura retórica, permitindo, assim, uma seleção "não cega", ao contrário do que ocorre com o Topline. Desse modo, para cada texto-fonte do córpus temos três sumários gerados.

Nas tabelas que evidenciam os dados aqui apresentados, indicamos o número de cadeias de co-referência existentes em cada sumário (#CCR#) e o número de quebras de CCR verificadas nestas cadeias (#quebras#). Na contagem das CCR foram incluídas tanto as cadeias íntegras quanto as que apresentaram quebra.

Verificou-se que uma parcela considerável dos sumários gerados é idêntica para os três sistemas, ou seja, são formados pelas mesmas palavras, na mesma posição. Do total de 47 textos-fonte, em 15 deles ocorre essa identidade nos sumários, conforme podemos ver na Tabela 8:

Tabela 8. Sumários iguais com quebras de CCR ou não

Texto	**RheSumaRST**	**Topline**	**Saliência**
	#quebras	**#quebras**	**#quebras**
1. cotidiano1	0	0	0
2. cotidiano3	0	0	0
3. cotidiano4	0	0	0
4. cotidiano6	0	0	0
5. cotidiano7	0	0	0
6. cotidiano8	0	0	0
7. cotidiano9	0	0	0
8. cotidiano10	1	1	1
9. ciência1	0	0	0
10. ciência2	0	0	0
11. ciência3	0	0	0
12. ciência5	0	0	0
13. ciência6	0	0	0
14. informática5	1	1	1
15. mundo1	0	0	0
TOTAL	**2**	**2**	**2**

Para outros 28 textos, cada conjunto de sumários apresenta sumários idênticos entre dois dos três sistemas, e apenas para quatro textos temos sumários distintos para cada um dos sumarizadores (Tabela 9).

Tabela 9. Sumários distintos entre os sistemas de sumarização

			RHeSumaRST		Topline		Saliência	
texto	**#markables**	**#CCR**	**#quebras**	**#CCR**	**#quebras**	**#CCR**	**#quebras**	**#CCR**
Esporte1	**26**		**0**	**3**	**1**	**2**	**0**	**1**
Esporte4	**24**	**4**	**0**	**1**	**0**	**1**	**1**	**1**
Informática10	**12**	**4**	**0**	**2**	**2**	**2**	**0**	**1**
Mundo10	**11**	**2**	**1**	**1**	**0**	**0**	**0**	**0**

Considerando, porém, o universo de 47 textos, independentemente da produção de sumários distintos ou não, obtivemos os seguintes resultados, com relação ao número de ocorrências de quebras de CCR (Tabela 10):

Tabela 10. Cômputo geral de quebras

RHeSumaRST		**Topline (baseline)**		**Saliencia**	
#quebras	**#CCR**	**#quebras**	**#CCR**	**#quebras**	**#CCR**
13	**89**	**16**	**83**	**8**	**60**

Como podemos observar, os resultados obtidos pelo RHeSumaRST são semelhantes aos do *baseline* – RHeSumaRST => 14% (89/13); Topline => 19% (83/16); e Modelo de Saliência => 13% (60/8). Estes dados evidenciam que o RHeSuma-2 possui fragilidades e que estas podem estar relacionadas a problemas de segmentação e de domínio dos textos escolhidos.

Estes resultados preliminares apenas apontam para os índices de quebras de CCR verificados nos sumários produzidos. Persiste, porém, a necessidade de uma avaliação mais acurada acerca da natureza destas quebras, ou seja, a investigação das causas. Para o mesmo experimento descrito nesta seção, a análise mais detida demonstrou que nem todos os casos computados como quebra podem ser atribuídos ao sumarizador em si, havendo problemas causados pela má segmentação do texto-fonte pelo DiZer e também problemas de anotação que induziram erros na etapa de avaliação dos sumários.

Tabela 11. Problemas verificados

#tipo de problema verificado	#número de casos
Segmentação do DiZer	1
Anotação do córpus	6
Desempenho do sumarizador	13

A partir dos dados apresentados, pudemos identificar o foco dos problemas causadores de quebras de CCR, fornecendo subsídios aos cientistas da computação envolvidos na tarefa de construção/programação dos sistemas de SA a fim de que melhorias fossem implementadas na arquitetura dos sistemas computacionais. Através da análise dos sumários automaticamente produzidos e da identificação dos casos em que a quebra é introduzida por parte do sistema de sumarização, foi possível direcionar a pesquisa ao foco do problema (tratamento computacional das CCR) e, assim, contribuir para a melhoria do desempenho desses sistemas.

8.1.1 Análise dos problemas

Apresentamos aqui exemplos dos problemas identificados para o córpus e indicados na Tabela 11. O relatório completo de todos os casos, analisados como segue nos exemplos dessa seção, pode ser encontrado em Carbonel & Rino (2006).

8.1.1.1 Problema de segmentação do DiZer

texto ciência_8: neste exemplo há um problema que decorre da própria segmentação do texto. O sumário do RHeSumaRST contém o markable 9, (os comentários), que aponta para o markable 11 (o primeiro-ministro britânico, Tony Blair, admitiu que o Protocolo de Kioto não está funcionando e pediu um novo acordo internacional para combater o aquecimento global). O sumário, porém, inclui esse markable incompleto – *o primeiro-ministro britânico, Tony Blair, admitiu que o Protocolo de Kioto não está* – o que é insuficiente para a compreensão do texto. Em vista disso, foi computada uma quebra;

O primeiro-ministro britânico, Tony Blair, admitiu que o protocolo de Kyoto não está funcionando e pediu um novo acordo internacional para combater o aquecimento global. **(markable 11)**
Em artigo publicado neste domingo pelo jornal The Observer, de Londres, Blair disse que, para que um novo acordo funcione, precisa incluir os Estados Unidos, que são os maiores emissores de gases poluentes do mundo, mas optaram por não ratificar o protocolo de Kyoto.
Os comentários **(markable 9)** foram feitos antes de uma conferência sobre mudanças climáticas em Londres na terça-feira.

Texto 1 – ciência 8 (texto-fonte)

O primeiro-ministro britânico, Tony Blair, admitiu que o protocolo de Kyoto não está **[...?...]** Em artigo publicado neste domingo pelo jornal The Observer, Os comentários foram feitos antes de uma conferência sobre mudanças climáticas em Londres na terça-feira.

Texto 2 – ciência 8 (sumário produzido pelo RHeSumaRST)

O sumário produzido pelo Modelo de Saliência, por sua vez, contém o referido markable 11 também "mutilado".

O primeiro-ministro britânico, Tony Blair, admitiu que o protocolo de Kyoto não está funcionando e pediu um novo acordo internacional Os comentários foram feitos antes de uma conferência sobre mudanças climáticas em Londres na terça-feira.

Texto 3 – ciência 8 (sumário produzido pelo Modelo de Saliência)

O markable completo seria:

O primeiro-ministro britânico, Tony Blair, admitiu que o protocolo de Kyoto não está funcionando e pediu um novo acordo internacional para combater o aquecimento global.

O segmento que foi extraído pelo Modelo de Saliência:

O primeiro-ministro britânico, Tony Blair, admitiu que o protocolo de Kyoto não está funcionando e pediu um novo acordo internacional

Notemos que ao indexar o markable 9 (os comentários) com o segmento extraído pelo MS, o sumário recupera a informação principal contida no markable antecedente. O mesmo não ocorre no sumário produzido pelo RS, no qual o segmento extraído não permite que o markable 9 recupere a informação, gerando, assim, uma inconsistência na leitura.

8.1.1.2 Problemas na anotação do córpus

texto ciência_4: nesse texto ocorre um problema no processo de anotação das CCR. O markable 6 (O estudo, que será publicado no British Medical Journal), aponta para o referente indicado como markable 19, (estudo). Os sumários resultantes do RHeSumaRST e do Topline incluem o markable 6 sem incluir o 19, o que por si só configura uma quebra. Todavia, do modo como está feita essa anotação, a referência parece conter mais informação relevante que o referente, o que não é verdade se considerarmos que, no texto-fonte, o estudo indicado pelo markable 19 continha mais informação que não foi incluída – foi computada uma quebra; o mesmo para o Modelo de Saliência.

Homens e mulheres de 50 anos ou mais deveriam tomar uma aspirina por dia para reduzir o risco de ataques cardíacos e derrames, segundo estudo **(markable 19)** realizado pelo médico Peter Elwood, da Universidade de Cardiff, no Reino Unido.
O estudo, que será publicado no "British Medical Journal" (**markable 6)**, avaliou 2.500 homens por 25 anos e concluiu que as pessoas de meia idade têm risco alto o suficiente para se beneficiar do uso do medicamento. Elwood defende a ampliação do uso da aspirina.

Texto 4 – ciência 4 (texto-fonte)

Homens e mulheres de 50 anos ou mais deveriam tomar uma aspirina por dia para reduzir o risco de ataques cardíacos e derrames,O estudo, que será publicado no British Medical Journal,

Texto 5 – ciência 4 (sumário produzido pelo RHeSumaRST e pelo Topline)

O problema, neste caso, é com relação ao conteúdo do markable 19, que 19 deveria ser: "O estudo realizado pelo médico Peter Elwood, da Universidade de Cardiff, no Reino Unido". Desse modo, contendo a informação completa, o referido markable pode ser indicado como antecedente do markable 6.

8.1.1.3 Problemas de desempenho do Sumarizador Automático

texto informática_10: o sumário produzido pelo Topline contém duas quebras. Há, no sumário, o markable 5 (o resort ainda sem nome do jogo que permite a interação entre milhares de usuários), mas não o 10 (um resort espacial construído no jogo on-line de RPG role-playing game Project Entropia), seu referente; e contém o markable 9 (os jogadores), mas não o 12 (milhares de usuários), seu referente – foram computadas duas quebras.

8.1.2 Considerações acerca do experimento

A análise dos sumários gerados no contexto do experimento descrito forneceu importantes informações acerca do fenômeno da referenciação, em especial a reformulação do protocolo de anotação das CCR no córpus de pesquisa. A perda da referência, como vimos, implica a incoerência dos sumários, prejudicando, assim, a utilidade do texto (informar acerca do conteúdo principal do texto-fonte). Em alguns casos, porém, os problemas de quebras de CCR não estão diretamente ligados ao processamento do RHeSumaRST, mas sim ao desempenho geral do DiZer.

O DiZer, enquanto um parser discursivo para o português, foi treinado e preparado para o processamento de textos do domínio científico, e não do jornalístico. Outra questão acerca do DiZer é a segmentação. Foram verificados casos em que a segmentação do texto-fonte introduziu um problema diretamente relacionado à perda da referência no sumário gerado. Em função desses problemas, portanto, era esperado que os resultados do DiZer fossem deficientes.

As observações acerca do experimento, por outro lado, apontam pontos fracos que, uma vez sanados, permitirão um melhor desempenho do sistema. As considerações sobre as CCR e a anotação das mesmas auxiliaram na elaboração de um novo protocolo de anotação (vide seção 5), permitindo uma análise mais ampla da construção da referência e de sua manutenção nos sistemas de SA – o problema das referências que extravasam o sintagma (encapsulation), por exemplo, foi um ponto estudado em nossas considerações. A utilização dessas modificações é reportada na próxima seção, em que descrevemos o Experimento 02.

8.2 Experimento 02 – Validação da Teoria das Veias e proposta de melhoria no modelo de SA utilizado

A Teoria das Veias foi utilizada inicialmente pelo RheSumaRST, como vimos na seção 8.2, com a finalidade de determinar o conjunto de possíveis termos antecedentes, possibilitando ao sistema a aplicação de heurísticas de sumarização mais específicas no que se refere à manutenção dos elos co-referenciais. Os experimentos envolvendo o sumarizador, no entanto, não contemplaram o algoritmo decorrente da Teoria das Veias, reportando apenas

os dados referentes à qualidade dos sumários gerados. Isso implica afirmarmos que, a despeito da aplicação da teoria em sistemas envolvendo a Língua Portuguesa, não encontramos na literatura qualquer menção a testes que a validassem para outras línguas que não aquelas eleitas para os testes feitos pelos próprios autores (inglês, francês e romeno). Os autores, no trabalho de 1998, não são explícitos quanto a questão da dependência de língua para a teoria. Podemos supor, porém, pelos testes realizados com língua diferentes (com desempenhos sutilmente diferentes), que não se trata de uma teoria independente de língua.

Seria possível partirmos da hipótese de que se os resultados obtidos pelos autores para essas três línguas diferentes (duas de origem românica e uma do tronco anglo-saxão) são aproximados entre si (tabela 12), então, para o português (uma língua também de origem românica) a precisão do algoritmo deveria ser, da mesma forma, um resultado aproximado aos obtidos no experimento original para aquelas três língua (acima dos 90%).

A discussão que propomos neste experimento (E2) visou não apenas demonstrar que é necessária uma validação da teoria para a Língua Portuguesa, mas apresentar valores que sejam, de fato, realistas, e que reflitam o verdadeiro poder preditivo da Teoria das Veias no que se refere à identificação dos candidatos a termos antecedentes em caso de relações anafóricas.

Apresentamos nesta seção, portanto, o cenário do experimento original, reportado por Cristea et al. (1998), a fim de estabelecer uma comparação com as características e resultados em E2 (Carbonel et al., 2007). Apresentamos, também, a ferramenta computacional desenvolvida para a realização deste experimento – o VeinTracker. cujas funcionalidades possibilitaram, no contexto deste experimento, a verificação automática da precisão nos cálculos do domínio de acessibilidade referencial (*acc*) para o textos do córpus utilizado.

8.2.1 Cenário do experimento original – Cristea et al., 1998

Conforme vimos na seção 5.3, a Teoria das Veias determina que, dado um texto *T*, para uma EDU *X*, existe um conjunto *V* (veia) que contém as outras unidades discursivas que

possam ter relação referencial com *X*. Vimos também que desse conjunto *V* é possível extrair o subconjunto *acc* (domínio de acessibilidade referencial), que contém apenas as proposições anteriores a *X*, ou seja, as proposições com contêm os possíveis candidatos a termo antecedente se *X* contiver uma expressão anafórica.

Para um experimento de avaliação da VT, portanto, é fundamental que seja utilizado um córpus com algumas características específicas: i) anotação de estruturas retórico-discursivas (RST) para os textos; ii) anotação das cadeias de co-referência. Por conta destas restrições sobre o córpus, Cristea et al., que tinham a proposta de verificar a teoria para três línguas distintas, utilizaram córpus bastante restritos e, no que se refere ao gênero, diferentes entre si.

Para o inglês, foram utilizados três textos curtos[47], anotados pelos próprios autores; para o francês foi utilizado um fragmento do romance *Pai Goriot*, de Honoré de Balzac, previamente anotado com cadeias de co-referência (Bruneseaux and Romary, 1997) e com anotação RST feita pelos autores; e, para o romeno, foi utilizado um fragmento da obra *As lendas do Olimpo* (história romanceada), de Alexandru Mitru, com todas as anotações feitas por um dos autores do trabalho. Como podemos ver, temos neste trabalho alguns pontos criticáveis na metodologia do experimento: i) não há uniformidade de gênero nos textos que compõem os córpus para as três línguas; ii) não há uniformidade nas anotações que marcam o córpus – os textos em inglês foram anotados pelos autores, o em francês já estava parcialmente anotado (co-referência) e os autores apenas anotaram as estruturas retóricas, e para o romeno toda a anotação foi feita apenas por um dos autores; iii) as descrições fornecidas prejudicam a reprodutibilidade do experimento, pois não é possível, a partir do artigo, identificar informações importantes, como que tipo de fenômeno co-referencial foi anotado, se o foco foi em um tipo específico de forma referencial, ou em todos os tipos etc. Quanto à dimensão dos textos utilizados, sabemos apenas o número de EDUs (Tabela 12), o que é muito vago, uma vez que não nos é informado o tipo de segmentação que foi realizada (se sentencial ou oracional).

[47] Os autores não dão qualquer outra informação sobre os textos, o que nos permite concluir, por oposição aos detalhes dados para os textos em francês e romeno, não serem literários.

Tabela 12. Verificação da Conjectura 1 (C1) da Teoria das Veias[48]

fonte	no. EDUs	Total exp.ref.	Referências Diretas (caso 1)		Referências Indiretas (caso 2)		Inferenciais (caso 3)		Quantos obedecem à C1	
inglês	62	97	75	77,3%	14	14,4	5	5,2%	94	96,9%
francês	48	110	98	89,1%	11	10,0%	1	0,9%	110	100,0%
romeno	66	111	104	93,7%	2	1,8%	5	4,5%	111	100,0%
Total	**176**	**318**	**277**	**87,1%**	**27**	**8,5%**	**11**	**3,5%**	**315**	**99,1%**

8.2.2 Cenário de E2: metodologia, resultados e discussão

Utilizamos um subconjunto do córpus Summ-it, composto por doze textos de divulgação científica do jornal Folha de São Paulo (caderno Ciência), totalizando 3.846 palavras, 430 EDUs e 1.156 expressões referenciais (ERs). Destas, testamos apenas os sintagmas nominais definidos, que totalizam 474 itens. Elegemos esse recorte por: i) contar com a anotação de cadeias de co-referência – análise feita por dois especialistas treinados (anotação descrita na seção 6); ii) contar com a anotação de estruturas retórico-discursivas – análise feita por dois especialistas treinados (anotação descrita na seção 7); iii) estando estas anotações coordenadas dentro de um mesmo projeto, haver uma preocupação com relação à consistência das anotações, o que se reflete pelo constante intercâmbio de informações e críticas entre os membros e colaboradores do projeto; iv) julgarmos que, usando essa importante categoria, realizaríamos uma validação preliminar que, além de nos permitir aprimorar nossa metodologia, também produzisse resultados de significância considerável.

A arquitetura do processo de validação pode ser observado na Figura 26(Carbonel et al., 2007).

[48] A tabela foi transcrita integralmente de Cristea et al. (1998), e foi traduzida visando facilitar a leitura.

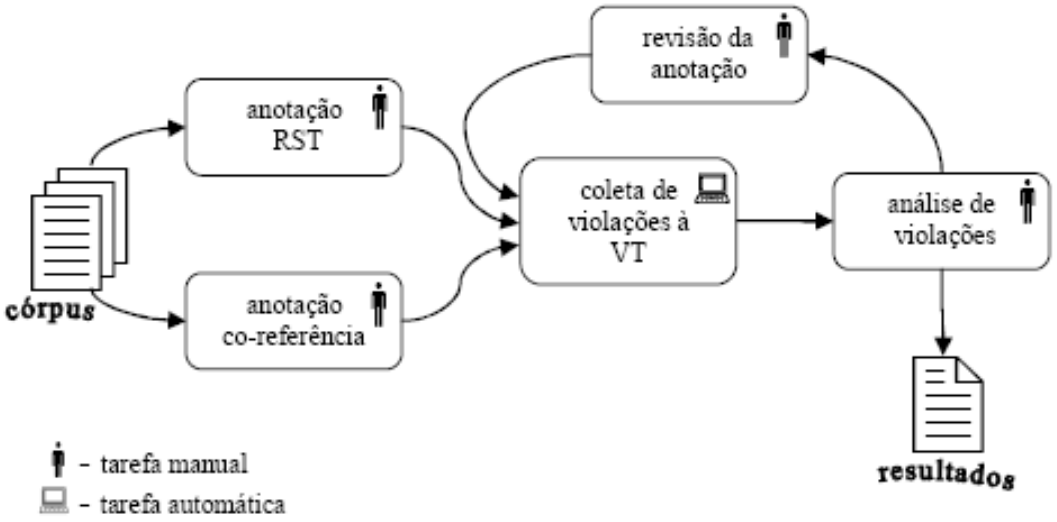


Figura 26. Processo adotado na validação da Teoria das Veias

Os textos foram processados com o algoritmo da VT, que calculou, para as estruturas RST dos mesmos, a veia e o domínio de acessibilidade referencial de cada EDU. O sistema de verificação automática – VeinTracker –, desenvolvido para este experimento[49], então, utiliza a anotação de co-referência para verificar se, nos casos de EDUs contendo expressões anafóricas, a EDU que contém o antecedente se encontra no *acc* da EDU referencial.

Na primeira etapa do experimento, o sistema retorna todos os casos em que a EDU contendo o antecedente não se encontra no *acc* da EDU anafórica. Sobre esses dados, é feita uma revisão manual que visa identificar um entre os seguintes problemas: i) anotação RST (quando houve um erro na anotação, do tipo inversão na relação núcleo-satélite, ou mesmo um problema de segmentação); ii) anotação de cadeias de co-referência (quando uma expressão nova no discurso é marcada como velha, por exemplo); iii) expressão anafórica cujo *acc* não contém a EDU na qual está o antecedente, mas cujo sentido é possível inferir a partir do contexto textual; e iv) uma violação real da VT (que ocorre quando não se tratar de nenhum dos casos anteriores).

[49] Projeto e Desenvolvimento em parceria com um cientista da computação – agradecemos a Jorge Marques Pelizzoni pela autoria do sistema.

Para os problemas descritos e (i) e (ii) identificados na primeira etapa, foram feitas modificações solucionando os mesmos e os textos foram novamente processados pelo algoritmo da VT e a anotação gerada submetida ao VeinTracker, de modo que fosse possível verificar se os erros eram, de fato, introduzidos por problemas de anotação dos textos – é o que podemos observar na descrição feita na Figura 26. Os resultados depurados por essa revisão manual podem ser vistos na Tabela 13 (acompanhados dos resultados reportados por Cristea et al. em seu trabalho, de modo a permitir uma melhor comparação entre nossos resultados), onde ERs são as expressões referenciais e os algarismos romanos indicam os casos da Conjectura 1 (C1) da Teoria das Veias, apresentados na seção 4.3. Usamos a notação I $\cup$ II, pois Cristea et al., no referido trabalho, não apresentam os casos separadamente.

Tabela 13. Resultados de verificação do cálculo das veias para os textos em português

Língua	EDUs	ERs	Testáveis	– I – ERNs	– II – Diretas	I ∪ II	– III – Indiretas	– IV – Infer.	C1-ok	Precisão PNT
inglês[50]	62	97	97	?	?	75 77,3%	14 14,4%	5 5,2%	94 96,9%	?
francês[50]	48	110	110	?	?	98 89,1%	11 10%	1 0,9%	110 100,0%	?
romeno[50]	66	111	111	?	?	104 93,7%	2 1,8%	5 4,5%	111 100,0%	?
port.	**430**	**1.156**	**474**	**301** **63.5%**	**108** **22.8%**	**409** **86.3%**	**17** **3.6%**	**18** **3.8%**	**446** **94.1%**	**81.94%**

Dignas de explanação mais detalhada são as duas últimas colunas da tabela: C1-ok denota o conjunto de testáveis que estão em conformidade com a C1; e a última coluna se refere ao nosso cálculo particular de precisão, dito Precisão Não-Trivial (PNT), dado pela seguinte fórmula:

$$\mathrm{PNT} = 1 - \frac{\left|\overline{\mathrm{C1\text{-}ok}}\right|}{\left|\overline{\mathrm{I} \cup \mathrm{IV}}\right|},$$

[50] Dados reproduzidos de Cristea et al. (1998) para fins de comparação.

onde $|X|$ e $\overline{X}$ denotam respectivamente o número de elementos do conjunto X e o complemento de X em relação ao universo de testáveis. Intuitivamente, PNT é o complemento de uma taxa de erro mais realista, dada pela razão entre o número de erros – $|\overline{\text{C1-ok}}|$ – e o total de casos não-triviais – $|\overline{\text{C1-ok}}|$ – cobertos ou não pela C1. Esse cálculo se justifica por sabermos que nenhum dos erros em $\overline{\text{C1-ok}}$ jamais deveria pertencer aos casos I ou IV.

Nossos resultados diferem significativamente dos de Cristea et al. Primeiramente, a precisão daquele conjunto de experimentos é impressionante: apenas 3 ERs não cobertas em 318, todas para o inglês e corrigíveis pela simples conversão de uma relação hipotática em paratática (uma ATTRIBUTION mononuclear que é convertida para multinuclear[51]). Entretanto, mais notável ainda é o fato de que a proximidade lingüística do português com o francês e o romeno não fique aí refletida. Isso parece sugerir que diferenças lingüísticas não devem ser responsáveis pelo contraste observado. Antes, cremos que tenham sido determinantes diferenças relativas a (i) gêneros textuais no córpus e (ii) esquema de anotação RST.

Quanto a gêneros textuais, podemos afirmar que, pelo menos para o romeno e o francês, foram usados fragmentos de narrativa como córpus, um em cada caso; enquanto o gênero dos três textos em inglês não foi mencionado. De nossa parte, usamos textos de divulgação científica, sabidamente de estrutura retórica distinta dos narrativos. Entretanto, cabe lembrar que há carência de estudos mais aprofundados sobre o impacto do gênero sobre a estrutura RST, especialmente aqueles que se atenham à topologia das árvores e à distribuição de nuclearidade, como a VT.

Quanto a esquemas de anotação RST, sabe-se que há grande variação tanto em tipologia de relações quanto diretrizes de segmentação (em EDUs) e anotação propriamente dita. Infelizmente, Cristea et al. não especificam qual esquema usam.

[51] O conjunto de Marcu (1997) prevê a possibilidade de termos ATTRIBUTION mono e multinuclear.

Finalmente, a inovação de usar a PNT se prova válida ao desfazer qualquer otimismo com a precisão geral de 94,1% observada. Um poder preditivo real de 81,94% certamente comporta melhorias e revela uma C1 bem menos absoluta que aquela reportada no trabalho de Cristea et al., inicialmente. Diante desses dados, portanto, cumpre analisar os casos problemáticos a fim de identificar as causas (que em alguns casos consistem em limitações da própria língua) dos erros no cálculos das veias e do domínio de acessibilidade referencial.

Em primeiro lugar, é importante ressaltar que toda tarefa de anotação de córpus está sujeita a erros, e todo esquema de anotação faz um recorte fenomenológico que pode se provar insuficiente em certas situações, para não mencionar a própria possibilidade de conter erros conceituais. Durante a depuração dos nossos resultados, encontramos todas essas situações em ambas as modalidades de anotação utilizadas.

RST. Quanto à anotação RST, contamos seis erros comuns de anotação, ora de segmentação, ora de inversão de nuclearidade (trocar núcleo por satélite e vice-versa). Houve mais 7 casos de falha devido à relação *attribution*, muito comum no gênero jornalístico. Trata-se de um caso que chegamos a considerar como uma falha conceitual no nosso esquema de anotação. Segundo Carlson & Marcu, na relação *attribution* (hipotática), o núcleo apresenta a expressão, fala ou pensamento de alguém, ao passo que o satélite indica o respectivo emissor. No exemplo E1, apresentamos um caso identificado no córpus.

E1: ["Em oito anos, detectamos mais de 300 eventos, graças ao nosso sistema de calibragem dos dados de satélite"]$_N$ [, *conta Douglas Revelle, do Laboratório Nacional de Los Alamos, um dos autores do estudo, que está publicado na edição de hoje da revista britânica "Nature" (www.nature.com).*]$_S$

Observamos que, no texto jornalístico, é extremamente comum a introdução de novos referentes no satélite de relações *attribution*. Além disso, é usual que esses referentes sejam retomados posteriormente no discurso. Entretanto, é um corolário da VT que jamais um satélite *S* numa subárvore de raiz *R* pertence ao *acc* de qualquer nó cujo caminho para *S* passe por nós acima de *R*. Assim, é muito comum que as referências posteriores não satisfaçam a C1. Esse problema pode ser evitado se considerarmos todas as relações

attribution como paratáticas (multinucleares). Isso permitirá o acesso a *S* pelo menos pela subárvore de todo ancestral *R*' de *R* tal que só haja arestas N separando *R* de *R*'.

Co-referência. Quanto à anotação de co-referência, houve três casos de erro trivial de anotação, de marcação de ER nova no discurso com anafórica. Por outro lado, no que concerne a deficiências conceituais, o quadro é um pouco mais complexo do que para RST. Em específico, nosso esquema se concentra numa co-referência estrita, sem explicitar a possibilidade de resolução de certas ERs na ausência de suas ERs co-referentes precedentes, ou seja, a (in)dependência de uma ER em relação às demais para ser interpretada. Referências inferenciais (caso IV da C1) constituem o caso mais freqüente dessa situação. Veja os exemplos seguintes[52]:

E2: "... [o País]$_{i,nova}$ ... [o Brasil]$_{i,velha}$ ..."

E3: "... [o fígado]$_{j,nova}$ ... [células de [o fígado]$_{j,velha}$]$_{k,nova}$ ..."

Nos exemplos E2 e E3, temos as ERs "o Brasil" e "o fígado" (2ª ocorrência), claramente interpretáveis na ausência de seus antecedentes, marcadas como ERs anafóricas quaisquer pela simples razão de nosso esquema de anotação não distinguir esses casos.

Em termos de revisão do esquema de anotação (e não de custo de anotação ou mesmo reprodutibilidade desta por computador), a solução para os casos de ERs inferenciais é trivial, podendo ser efetuada pela mera adição de um traço de anotação. Entretanto, existem diversos outros exemplos menos claros, que fogem ao escopo das soluções implementadas neste experimento por denotarem uma complexidade lingüística atávica à própria língua:

E4: "[Um ser que invade corpos e domina a mente alheia, forçando suas vítimas a fazer o que ele ordena,]$_{i,nova}$ não é mero personagem de ficção. Para uma aranha da Costa Rica, essa criatura existe ... Apesar do nome *Hymenoepmecis sp.*, [o tal invasor de corpos]$_{i,velha}$ é só [uma vespa]$_{i,velha}$ ..."

[52] Para recuperar contexto do exemplo, ver texto CIENCIA_2000_17109 no apêndice B.

Em E4, a ER “uma vespa” é anotada como anafórica por sua co-referência com ERs anteriores, as quais se encontram em satélites que não estão acessíveis a ERs posteriores verdadeiramente dependentes de “uma vespa”. Estas ERs são consideradas, então, como violações à C1, apesar de acessarem a EDU onde se encontra “uma vespa”.

Aos lermos o texto inteiro, observamos um fenômeno interessante na construção da referência: nas quatro primeiras menções à entidade “vespa”, o leitor não sabe ainda tratar-se de uma vespa. Tem-se, na verdade, uma preparação do leitor (que utiliza uma espécie de “suspense retórico”, à medida em que permite que o mesmo estabeleça relações com o universo ficcional, o que, aliás, é explicitado no texto), para, apenas na quinta menção, nomear-se a entidade, esclarecendo que o objeto de discussão é uma vespa. Segundo os guidelines de anotação co-referencial, essa cadeias foi anotada da maneira a definir na primeira linha o termo referente (menção inaugural no discurso) e nas demais as expressões referenciais, conforme apresentado na Figura 27.

Classificação		**Sintagma**
CADEIA : set_33		
word_1..word_19	---	**Um ser que invade corpos e domina a mente alheia , forçando suas vítimas a fazer o_que ele ordena**
word_18	---	**ele**
word_34..word_35	---	**essa criatura**
word_49..word_56	---	**-Hymenoepimecis sp .- o tal invasor de corpos**
word_59..word_60	---	**uma vespa**
word_75..word_76	---	**esse inseto**
word_114..word_115	indirect--old	**a vespa**
word_145..word_146	indirect--old	**a Hymenoepimecis**
word_235..word_236	---new	**a parasita**
word_266	---	**ela**
word_322	---	**parasita**

Figura 27. Representação de uma das cadeias de co-referência do texto CIENCIA_2000_17108

Na marcação das veias e do domínio de acessibilidade referencial deste texto, a expressão referencial “uma vespa” não acessa diretamente “Um ser que invade corpos e domina a

mente alheia, forçando suas vítimas a fazer o que ele ordena", seu antecedente. As expressões posteriores também não acessam este antecedente, mas acessam "uma vespa" e tal acessibilidade corresponde às necessidade interpretativas do leitor, pois imaginemos que, hipoteticamente, a anáfora pronominal "ela" (word_266) acessasse "Um ser que invade corpos e domina a mente alheia , forçando suas vítimas a fazer o que ele ordena", sem acessar "uma vespa". Nesse caso, faltariam informações ao leitor para depreender o que, exatamente, o pronome "ela" recupera.

O que temos nesse exemplo é, aparentemente, um caso de progressão referencial, no qual o sentido é construído no decorrer do texto e só se estabiliza em um determinado ponto – o que ocorre apenas nas word 114 e 115, ou seja, após uma razoável porção de texto se considerarmos que é o assunto central do mesmo. Isso, então, evidencia uma característica interessante (e particularmente complexa do ponto de vista do processamento computacional) da língua: a construção da referência.

Considerando, portanto, que a referência é apenas um encadeamento de expressões que remetem exatamente a um elemento inicialmente introduzido no discurso, mas sim uma progressão de sentidos que só se estabilizam em certo ponto do texto, a resolução anafórica (humana ou automática) consiste em um desafio maior do que apenas identificar o antecedente. Nesse sentido, podemos afirmar que uma anáfora só é efetivamente resolvida se é permitido ao leitor acessar o termo referente estabilizado no discurso, ou seja, o real antecedente cujo conteúdo semântico pretendeu ser recuperado pelo autor no ato da escrita.

E5: "... poderiam originar [as células hepáticas]$_{j,velha}$, além de [as sangüíneas]$_{k,velha}$"

Temos em E5, um caso curioso, em que a ER "as sangüíneas" não é co-referente com "as células hepáticas", mas depende desta para ser interpretada. Esse tipo de dependência não é capturado por nosso presente esquema de anotação.

Feitas estas considerações sobre a análise da Teoria das Veias para o Português, bem como a avaliação dos sistemas apresentada nesta seção, passamos, na seção 9, aos desdobramentos principais deste trabalho e nas propostas que, efetivamente, derivamos do trabalho apresentado até o presente momento.

9. Estudo e validação de teorias do domínio lingüístico com vistas à melhoria do tratamento de cadeias de co-referência em Sumarização Automática

Conforme vimos nas seções anteriores, o modelo de SA que estudamos envolve a aplicação conjunta da Teoria das Veias e da RST, baseando-se ainda em um modelo de saliência (Marcu, 1997) para a seleção da informação mais relevante – e o RheSumaRST é a implementação deste modelo que usamos como parâmetro, principalmente no Experimento 01. Neste capítulo, após havermos explorado detalhadamente as teorias aplicadas ao modelo, apresentamos uma avaliação do modelo em si, da qual derivamos uma proposta de reimplementação do mesmo, inserindo modificações com base nas críticas feitas.

9.1 Crítica ao RheSumaRST

Recuperando, em breve síntese, o que vimos na seção 8.2, o RheSumaRST aplica as teorias mencionadas acima, de modo a permitir a poda, para a qual aplica um critério de seleção que visa à manutenção dos elos referenciais. De acordo com este critério, o sistema prioriza a inclusão de todas as EDUs contidas nas veias das EDUs inicialmente selecionadas pelo modelo de saliência aplicado. Assim, tomemos um exemplo hipotético de um texto composto por 10 EDUs, classificadas (por sua saliência) do seguinte modo:

[+ saliente] EDU 1 > EDU 3, EDU 4 > EDU 2 > EDU 5, EDU 6 > EDU 8, EDU 10 > EDU 9 [- saliente]

Ao aplicarmos uma taxa de compressão de 30% (esperando obter sumários com cerca de 30% do tamanho do texto-fonte), e considerando que as EDUs tenham uma mesma extensão interna no que se refere ao número de palavras, teríamos um sumário formado pelas EDUs 1, 3 e 4. Como o sistema objetiva a manutenção das cadeias de co-referência, as veias (e não apenas o domínio de acessibilidade referencial) serão consideradas no processo de seleção:

EDU 1 (veia = EDU 1), EDU 3 (veia = EDU 1, EDU 3), EDU 4 (veia = **EDU 2**, EDU 4, **EDU 6**)

O sistema identifica, portanto, que, para manter os elos referenciais, é necessário inserir, além das EDUs inicialmente escolhidas, as EDUs 2 e 6, a fim de manter a coesão

referencial do sumário, que agora é composto pelas EDUs 1, 2, 3, 4 e 6. Como foram inseridas novas EDUs, o sistema opera recursivamente o mesmo processo anterior, verificando as veias dos novos componentes do sumário:

EDU 1 (veia = EDU 1), EDU 2 (veia = EDU 1, EDU 2), EDU 3 (veia = EDU 1, EDU 3), EDU 4 (veia = EDU 2, EDU 4, EDU 6), EDU 6 (veia = EDU 1, **EDU 5**, EDU 6)

Ao considerar a veia da EDU 6, o sistema ainda inseriu a EDU 5, passando o sumário a ser composto pelas EDUs 1, 2, 3, 4, 5 e 6. Aplica-se a recursão novamente para contemplar a veia da EDU 5, que contém as EDUs 1, 3 e 5. Como estas já compõem o sumário, o sistema pára e considera feita a seleção para a composição do sumário final. Como podemos observar neste exemplo meramente hipotético, há uma violação considerável da taxa de compressão – de 30% iniciais, terminamos com um sumário correspondente a 60% do texto-fonte.

Nos trabalhos de Seno (2005), não encontramos uma avaliação específica das violações da taxa de compressão. A premissa básica do referido sistema é preservar, a todo custo, todo contexto referencial possível, mesmo que isso implique transgredir o tamanho pré-estabelecido pelos parâmetros de compressão – em outras palavras, o RheSumaRST privilegia a possível coerência em detrimento da extensão dos sumários.

Realizando exatamente esse contexto experimental com o Córpus Summ-it, geramos sumários automáticos (com o RheSumaRST) com taxa de compressão de 30%, obtendo sumários com uma média de 40,63% do texto-fonte, o que corresponde a uma violação média de 35,43% com relação à proposta inicial de compressão, como mostra a Tabela 14.

Tabela 14. Análise da taxa de compressão do RheSumaRST

# text id #	# texto-fonte #	RheSuma extensão
2000_17082	270	111 (41,11%)
2000_17088	370	146 (39,45%)
2000_17101	315	127 (40,31%)
2000_17108	282	191 (67,73%)
2000_17109	241	89 (36,92%)
2000_17112	291	134 (46,04%)
2000_17113	368	147 (39,94%)
2002_22023	377	176 (46,68%)
2003_24219	360	167 (46,38%)
2004_26415	268	133 (48,50%)
2005_28747	288	130 (45,13%)
2005_28756	455	163 (35,82%)
TOTAL	**3385**	**1714**
MÉDIA	**324**	**143 (40,63%)**

A implementação de um sistema como o RheSumaRST obviamente lida com a possibilidade de variação da taxa de compressão nos sumários gerados – sumários menores ou maiores que o determinado pelo usuário do sistema. Para a taxa de 30%, sumários entre 25% e 35% de compressão com relação ao texto-fonte podem ser considerados aceitáveis[53]. Para os sumários gerados pelo RheSumaRST, apenas um sumário (2005_28756) se enquadra nesse parâmetro, com pouco menos de 36% de compressão. Ao observarmos o caso mais gritante – texto 2000_17108 – vemos que o sumário gerado neste texto viola a taxa de compressão em 125,8%, ou seja, é mais de duas vezes maior que o inicialmente proposto para o sumário, chegando, nesse caso, a não poder ser considerado uma versão condensada do texto-fonte.

[53] Parâmetro puramente empírico. Nos trabalhos da DUC, o parâmetro recomendado é exatamente a taxa de compressão; para este trabalho, adotamos uma tolerância maior, permitindo esta variação "além" da taxa máxima, de modo a não penalizar excessivamente nenhum dos dois sistemas.

Nossa proposta de melhoria, nesse sentido, ao mesmo tempo que objetiva a preservação da coerência, mais especificamente no tocante ao fenômeno do encadeamento referencial, considera importante não transgredir significativamente a taxa de compressão, uma vez que a meta básica desse tipo de sistema é a produção de sumários. No RheSumaRST, a transgressão se dá, sobretudo, pela condição de se inserir incondicionalmente toda veia completa de cada EDU que é escolhida para compor um sumário, como vimos no exemplo anterior.

Como vimos na seção 5.3, a veia de uma EDU abrange todos os elementos encadeáveis referencialmente à mesma, incluindo não apenas as anáforas, como também as catáforas. Ao processo de resolução anafórica – que é o pretendido pelo RheSumaRST – interessam apenas as EDUs candidatas a termo antecedente, ou seja, anteriores à EDU em questão. Para isso, a própria Teoria das Veias prevê o domínio de acessibilidade referencial.

Ao considerar a veia inteira, o RheSumaRST aumenta, portanto, consideravelmente o número de elementos inseridos, o que pode explicar as significativas violações à taxa de compressão.

Resta, por fim, reconsiderar uma das premissas essenciais do RheSumaRST à luz da validação das Teoria das Veias para o português. Conforme já dito anteriormente, ao utilizar a VT para aplicações em Língua Portuguesa, Seno não contava com dados específicos para a língua, baseando-se apenas nos resultados (quase absolutos, lembremos-nos) apresentados por Cristea et al. para outras línguas. Por conta disso, talvez, Seno assumiu em seu trabalho que a inserção das veias dos constituintes do sumário garante, apenas com uma pequena margem de erro, a manutenção dos elos referenciais anafóricos. Demonstramos neste trabalho, porém, que os tipos de resolução anafórica de fato relevante são aqueles que chamamos de não-triviais, para os quais a cobertura da VT é de aproximadamente 80%.

Sintetizando, os problemas que podemos identificar no RheSumaRST – considerando-se apenas os aspectos qualitativos dos resultados gerados pelo sistema[54] – destacamos: a violação excessiva da taxa de compressão e a utilização da veia inteira (e não apenas do domínio de acessibilidade referencial) como parâmetro para a garantia da manutenção dos elos.

Com base na identificação destes pontos problemáticos, propusemos uma reimplementação do RheSumaRST com base em alterações significativas em algumas etapas do processo. É o que passamos a descrever na próxima seção.

9.2 VeinSum: uma nova proposta de implementação para o RheSumaRST

Como vimos na seção anterior, os pontos críticos nos resultados gerados pelo RheSumaRST são um número ainda significativo de quebras de cadeias de co-referência aliado, obviamente, à violação da taxa de compressão . Os sumários apresentam alta taxa de violação da compressão estabelecida, quebras de CCR e informatividade regular (aquém do desejável em um sumário informativo)

Nossa proposta com o VeinSum[55] apresenta algumas modificações na organização da arquitetura do RheSumaRST e uma nova estratégia de seleção de EDUs para compor o sumário, ainda baseada no Modelo de Saliência de Marcu, mas com uma aplicação que objetiva, além da manutenção dos elos referenciais, a não transgressão da taxa de compressão, como podemos ver na tabela 15.

[54] Não consideramos, em nenhum momento, aspectos da arquitetura computacional do sistema.
[55] Implementação computacional de autoria de Jorge Marquez Pelizzoni.

tabela 15. Características do RheSumaRST e do VeinSum

	RheSumaRST	**VeinSum**
formato de entrada	SGML	XML
formato de saída	sumários em txt	sumários em XML com diversas informações pertinentes
tratamento das cadeias de co-referência	inserção do todo o contexto referencial	inserção do contexto referencial menos oneroso à taxa de compressão
taxa de compressão	corrompida em função da preservação do contexto referencial	mantida através de heurística que seleciona o contexto referencial menos oneroso
utilização da VT	utilização da veia inteira	utilização apenas do *acc*

O sistema tem como entrada um texto sob a forma de sua estrutura RST (anotação XML). Uma vez inserida no sistema, a estrutura é processada pelo algoritmo de saliência (MarcuRank) e pelo algoritmo de veias (AddVeins). Estas duas informações – ranking das EDUs contituintes do texto-fonte e *acc* de cada EDU – são os dados de entrada do RankSum, módulo do sistema que, respeitando a ordem de saliência fornecida pelo MarcuRank, irá aplicar a mesma heurística de poda proposta por Seno.

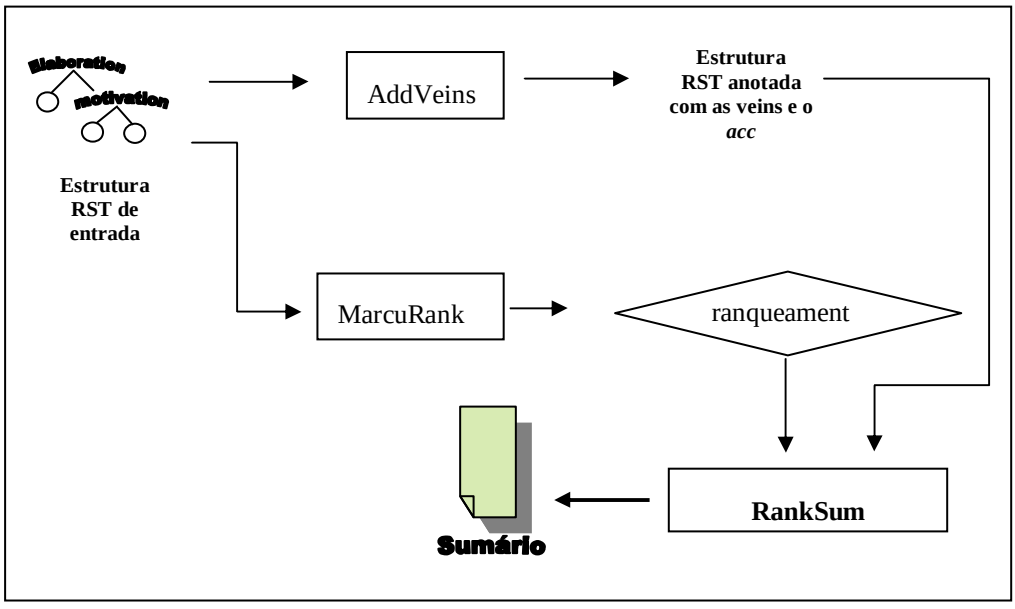


Figura 28. Arquitetura do VeinSum

Ao recuperarmos a arquitetura do RheSumaRST (seção 8), vemos que os dois sistemas compartilham alguns dos módulos descritos na Figura 28– o AddVeins, o MarcuRank e o Ranqueamento. A inovação da proposta do VeinSum, portanto, concentra-se na aplicação de uma abordagem distinta de seleção das EDUs ranqueadas, visando, além da manutenção dos elos referenciais, ao não corrompimento da taxa de compressão.

A proposta do VeinSum, no sentido de melhorar o desempenho do sistema de SA nesse aspecto pontual, consiste em descartar seqüências de EDUs mais relevantes que corrompem a taxa de compressão em função da inclusão de seqüências menores, porém menos relevantes. Nossa hipótese é que, ao incluirmos EDUs cujos *acc*s sejam menos onerosos à compressão do sumário, mesmo que estas sejam menos relevantes (de acordo com o modelo de saliência), estaremos construindo sumários menores, com menor inclusão de expressões referenciais e, por conseguinte, com menos quebras de CCR. O processo feito pelo RankSum pode ser visto na Figura 29.

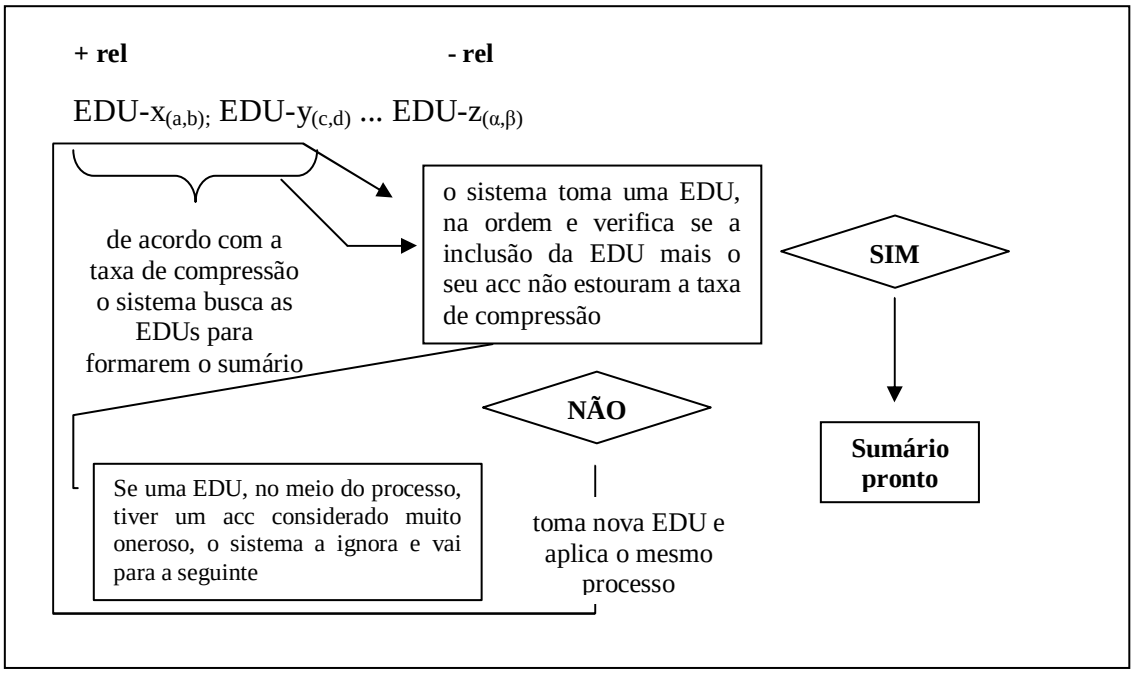


Figura 29. Arquitetura interna do RankSum

Nossa hipótese é a seguinte: adotando seqüências menos relevantes, porém menores, nós conseguiríamos resultados melhores no tocante à taxa de compressão e mantemos, ainda assim, o foco na informatividade do sumário, pois a EDU mais relevante (a melhor pontuada no ranqueamento de Marcu) seguramente seria incluída. Tomemos como exemplo um dos textos do Córpus Summ-it (vide apêndice B – CIENCIA_2005_28747). Reproduzimos na Figura 30 a anotação XML que o VeinSum fornece como saída.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<extract method="rank-prune/marcurank/ranksum"
  edus="[1,3,5,10,21,22,24,30,31]" spine="[3,5,22,24,10,31,30]"
  ignored="[16,8,7]"
  ranking="[3,5,22,24,10,31,7,8,16,30,1,9,11,15,25,27,28,2,6,13,14,19,20,
  29,4,12,17,23,18,21,26]" origlen="328" maxlen="98" len="104"
  maxrate="30.00" rate="31.71" delta="5.38">Chineses e americanos
  enxergam o mundo de jeitos distintos Pesquisadores da Universidade de
  Michigan em Ann Arbor, nos Estados Unidos, sugerem que os olhos de
  asiáticos tendem a ver uma imagem no seu conjunto enquanto os
  americanos demoram mais o olhar no objeto central de um quadro. Ele
  credita à sua colega Hannah Faye Chua a idéia de testar de forma visual
  um dado já verificado verbalmente. Para Nisbett, diferenças culturais
  explicariam essa assimetria. Nisbett e Chua pretendem agora ver se
  diferenças como essas se manifestam entre outras culturas. O estudo
  está na revista "PNAS".</extract>
```

Figura 30. Dado de saída do VeinSum para o texto CIENCIA_2005_28747

Essa anotação fornece um conjunto de informações que nos indicam o processamento feito internamente. Vemos o ranking de saliência das EDUs (ranking="[3,5,22,24,10,31,7,8,16,30,1,9,11,15,25,27,28,2,6,13,14,19,20,29,4,12,17,23,18, 21,26]") e as EDUs que são selecionadas (para compor o sumário) pela sua relevância (spine="[3,5,22,24,10,31,30]"). Para chegar a este conjunto *spine*, o sistema toma as EDUs na sua ordem de relevância, mas pode ignorar seqüências de EDUs que, por sua extensão, fariam estourar a taxa de compressão – é o que temos no conjunto *ignored*="[16,8,7]", preteridas pela EDU 30. É, então, a partir da *spine* que o sistema olha quais outras EDUs devem ser selecionadas por estarem no *acc* de alguma das já escolhidas – no caso do exemplo utilizado, o sistema selecionou as EDUs 1 e 21.

Como dados complementares, interessantes na avaliação dos resultados, o sistema ainda fornece outras informações, como:

- origlen: tamanho do texto-fonte, em número de tokens[56] (o número efetivo de palavras pode ser menor);
- maxlen: tamanho máximo que o sumário deve ter, considerando-se a taxa de compressão;
- len: tamanho do sumário gerado;
- maxrate: taxa de compressão que o sistema deve seguir;
- rate: taxa de compressão efetivamente alcançada;
- delta: variação ca compressão com relação à maxrate.

Para avaliar esse sistema, foram utilizados doze textos do Córpus Summ-it. Os resultados serão apresentados na próxima seção, antecedidos pelo detalhamento dos critérios de avaliação da informatividade.

9.3 Avaliação com base na coesão referencial

A contagem do número de quebras de cadeias de co-referência é uma medida de avaliação da qualidade textual dos sumários gerados automaticamente por indicar, pontualmente, os casos específicos em que, em tese, uma inconsistência estrutural insere um problema de ordem superficial (organização estrutural ou linear) que leva a uma deficiência na organização reticulada do texto – organização das idéias (estrutura profunda).

Essa consideração inicial é relevante porque leva em consideração um aspecto bastante importante do estudo do fenômeno co-referencial: a questão das quebras de cadeias de co-referência. Ressaltamos, também, que, apesar de darmos destaque para as descrições definidas em outras etapas de nossa pesquisa (avaliação da Teoria das Veias, por exemplo),

[56] Vide seção 2.1 para recuperar definição de *token.*

nesta etapa de avaliação do VeinSum utilizamos a marcação de CCR completa, incluindo, além das descrições definidas, as outras formais, tais como pronomes e expressões indefinidas.

Em nosso trabalho, não nos ocupamos do estudo pontual e analítico dessas quebras – por questões operacionais, tais como a não-disponibilidade de um córpus extenso e significativo que propiciasse, através de experimentos, casos de quebras em número e variedade suficientes para uma análise consistente. Usamos, na avaliação, critérios clássicos e já utilizados em experimentos anteriores (Carbonel et al., 2006), que ficam explicitados no esquema da Figura 31:

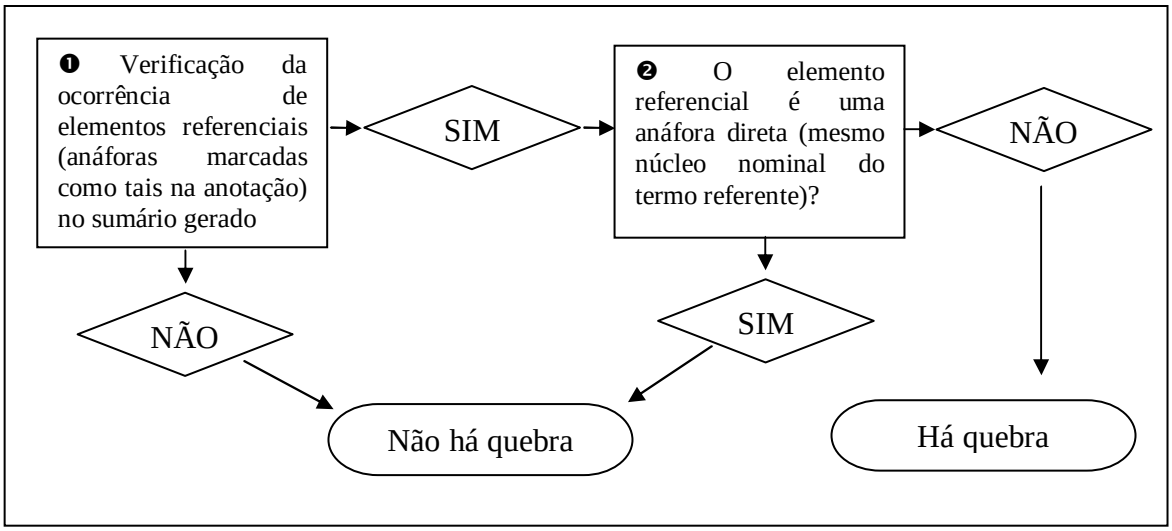


Figura 31. Avaliação dos casos de quebra de CCR

Esse tipo de avaliação, no entanto, se dá em um nível muito superficial de análise, pois não leva em consideração o tipo de relação de dependência há entre o elemento referencial e seu referente. Podemos, a partir de dois exemplos retirados de nosso córpus, ilustrar como dois casos de quebra são diferentes com relação ao impacto que exercem com relação ao contexto do sumário:

E_1 (CIENCIA_2000_17108): "não é mero personagem de ficção. Para uma aranha da Costa Rica, essa criatura existe. Apesar do nome o tal invasor de corpos é só uma vespa.(...)"

E_2 (CIENCIA_2000_17109): "O estudo, mostra que, além disso, elas são capazes de originar outro tipo de célula dentro do organismo humano."

Nos exemplos acima, temos três expressões marcadas, na anotação, como referenciais e seus antecedentes não se encontram nos sumários de onde as sentenças foram extraídas[57]: [essa criatura] em E_1 e [o estudo] e [ela] em E_2. No primeiro caso, notamos que, apesar na não inserção do termo antecedente no sumário, a expressão [essa criatura] sobre uma progressão na construção de sua referência, de modo que, através de referências catafóricas complementares – [o tal invasor de corpos] e [uma vespa] – o sentido é recuperado, mesmo sem o antecedente. Nos casos de E2, porém, as expressões [o estudo] e [ela], dependem exclusivamente do que vem antes (antecedente textual) para que o leitor tenha noção de a qual ente discursivo remetem. Ao analisarmos detidamente apenas as duas expressões, podemos ainda perceber que [o estudo] possui, com relação ao seu antecedente, uma relação de dependência menos significativa que [ela], pois, através das restrições semânticas inerentes à expressão, é possível limitar o escopo de busca no texto pelo termo antecedente (em outras palavras, o leitor tem um poder inferencial maior para fazer suposições acerca de "o que é" o estudo). O pronome pessoal [ela], porém, possui apenas restrições de número e gênero, aumentando, assim, as possibilidades interpretativas e dificultando o processo inferencial por parte do leitor.

Estes exemplos servem de base para a conclusão de que o fenômeno de quebra de CCR demanda um estudo mais acurado e dotado de um embasamento metodológico mais consistente que o contexto da presente pesquisa. Por esta razão mesmo, optamos por um olhar mais simplório que, dadas as propostas de aplicação de recursos lingüísticos visando à melhoria da resolução anafórica (o uso da Teoria das Veias, por exemplo), limitou-se à avaliação quantitativa, e não à qualitativa.

Analisando, então, o desempenho dos sumários automaticamente produzidos pelo VeinSum, tendo o RheSumaRST por baseline, chegamos à seguinte contagem de quebras, que podemos observar na tabela 16:

[57] Sumários e textos-fonte disponíveis nos apêndices B e C, ao final.

# text id #	VeinSum	RhesumaRST
	CCR qbr[2]	CCR qbr
2000_17082	1	3
2000_17088	0	1
2000_17101	0	1
2000_17108	0	3
2000_17109	0	1
2000_17112	0	0
2000_17113	0	0
2002_22023	1	1
2003_24219	0	4
2004_26415	0	2
2005_28747	2	0
2005_28756	0	4
TOTAL	**4**	**20**
MÉDIA	**0,33**	**1,66**

tabela 16. Avaliação das quebras de CCR

O interessante nos dados acima é que, apesar de adotarem, em linhas gerais, a mesma proposta metodológica de agregação da estruturação RST e o cálculo do domínio de acessibilidade referencial (Teoria das Veias), o desempenho dos dois sistemas foi bastante discrepante: o RheSumaRST apresentou 400% a mais de quebras que o VeinSum.

Acreditamos que as causas desses desempenhos distintos sejam de ordens diferentes e complementares. Em primeiro lugar, o RheSumaRST apresenta problemas estruturais relacionados à sua programação – os assim denominados "bugs"[58] – que podem ser melhor compreendidos através da leitura dos sumários produzidos pelo sistema[59]. No texto CIENCIA_200_17108, por exemplo (e este caso se repete em outros textos), a árvore RST de entrada possui uma relação, logo na primeira sentença, de SAME-UNIT, que é estrutural e multinuclear. O RheSumaRST, porém, separa as duas proposições e não insere a primeira.

[58] Essa avaliação não está relacionada ao nosso trabalho enquanto lingüistas, mas é um parecer fornecido pelos cientistas da computação que ampararam o desenvolvimento deste projeto.
[59] Vide apêndice C.

Além desses problemas, outra causa do número excessivo de quebras parece ser a grande quantidade de informação desnecessária que o sistema insere nos sumários, exatamente por não considerar um contexto referencial menos oneroso (aquele cujo acc não corrompe demasiadamente a taxa de compressão). Ao estudarmos os sumários produzidos pelo RheSumaRST, observamos sumários muito extensos e com muitas expressões anafóricas não resolvidas.

Essa constatação nos leva à hipótese de que a inserção de mais informação, em lugar de aumentar as possibilidades de inserção da informação contextual necessária à resolução anafórica, eleva o número de expressões referenciais a serem resolvidas. Desse modo, sumários menores, obtidos a partir de uma metodologia que visa a contenção, como é o caso do VeinSum, tendem a ter menos casos de quebras que sumários maiores.

9.3 Avaliação da Informatividade através da Medida ROUGE

A avaliação da informatividade dos sumários automáticos é de grande importância para a avaliação das propostas teóricas de um modelo de sumarização. Neste trabalho, além da medida tradicionalmente adotada na área (objetiva), calculada com o auxílio da ferramenta ROUGE (Lin, 2004a; Lin, 2004b)[60], utilizamos uma medida subjetiva, apresentada nesta seção.

Adotamos a medida ROUGE com uma dupla função: i) a de avaliar um sistema lingüístico-computacional através de uma medida reconhecida na área; e ii) validar, através de uma análise objetiva. Quanto à metodologia subjetiva, a adotamos por se tratar de uma abordagem mais rica em análise lingüística.

Na avaliação de informatividade dos sumários gerados automaticamente pelo RheSumaRST, Seno partiu da hipótese de Mani (2001), segundo a qual heurísticas baseadas na reprodução das informações constantes em sumários manuais garante a informatividade mínima dos sumários automáticos, uma vez que os sumários manuais são considerados ideais.

[60] Essa ferramenta foi adotada a partir da DUC de 2004 como apoio à avaliação automática (vide http://www-nlpir.nist.gov/projects/duc/data.html).

O parâmetro avaliativo apresentado pela autora é a comparação entre as ocorrências das relações retóricas nos sumários gerados manual e automaticamente, considerando a presença das mesmas relações nos dois sumários. Assim, se uma determinada relação ocorre, por exemplo, dez vezes nos sumários automáticos e se verifica sua ocorrência apenas cinco vezes nos sumários gerados, assume-se que a representatividade da relação é de 50%.

No contexto dos trabalhos em Sumarização Automática, uma ferramenta de avaliação bastante difundida é a ROUGE, que consiste em um pacote de medidas. Uma vantagem da ferramenta é a fácil reprodução desta avaliação e o baixo custo de se executá-la – se comparado com uma avaliação manual. Considera-se que a ROUGE possui a vantagem de ser uma ferramenta de avaliação dotada de consistência, evitando-se, assim, os erros humanos geralmente cometidos.

Baseada na medida BLEU (Pepineni et al., 2001), fortemente utilizada para a avaliação de sistemas de tradução automática, a ROUGE usa a abordagem de co-ocorrência de n-gramas, que consiste em verificar a média de quantas vezes cada conjunto de n palavras adjacentes se repetem em cada texto a ser avaliado.

O pacote da ROUGE utilizado oferece cinco medidas, tendo como elemento de comparação sumários de referência, considerados ideais porque produzidos manualmente:

- ROUGE-1: equivalente à medida unigramas, avalia a média do número de vezes uqe cada palavra aparece em cada um dos textos.

- ROUGE-2: mede bigramas, isto é verifica a freqüência de cada par de palavras que aparece em cada texto de entrada no sumário.

- A ROUGE-3 e ROUGE-4, semelhantes às medidas 3-grama e 4-grama, respectivamente, são pouco utilizadas, visto que a repetição de conjuntos de 3 ou 4 palavras adjacentes é muito incomum.

- ROUGE-L: Baseado na subseqüência comum mais longa (o *Longuest Common Subsequence* - LCS), busca as maiores sub-cadeias comuns entre os dois textos, executando então uma avaliação similar à co-ocorrência de n-gramas.

Em nossa pesquisa, utilizamos a ROUGE a fim de verificar a similaridade entre os sumários automáticos e os sumários de referência, o que permitiu analisar consistência de nossa avaliação lingüística, apresentada na seção anterior. Os dados que seguem correspondem à utilização da medida ROUGE-1, e foram produzidos para o VeinSum, para o RheSumaRST e para um baseline eleito para o experimento, o GistSumm (Pardo et al., 2002).

Como discutimos na subseção anterior, os sumários automáticos produzidos pelos diferentes sistemas de SA possuem, no mais das vezes, algum nível de corrupção da taxa de compressão, ora ultrapassando a taxa estabelecida (30% do texto-fonte), ora ficando aquém da mesma (o caso mais expressivo que temos é um sumário com 26,6% do texto-fonte). Por conta dessa diferença, particularmente entre o VeinSum e o RheSumaRST (sumários com média de compressão de 34,1% e 40,7%, respectivamente), o RheSumaRST apresentou uma medida de informatividade superior a do VeinSum. Ou seja, o RheSumaRST, por conter sumários com mais palavras, possui maior chance de, em comparação com os sumários manuais, apresentar melhores resultados. O VeinSum, por sua vez, alia dois objetivos: respeitar os parâmetros de compressão e manter os atributos textuais do sumário, particularmente a coesão referencial e a informatividade

Para minimizar, então, esta discrepância entre os sumários produzidos pelos sistemas, adotamos uma metodologia de uniformização dos sumários, baseada no truncamento do sumário maior. Este processo consiste simplesmente em cortar o sumário maior tendo como limite o número de palavras do sumário menor, tornando aquele comparável a este. Em nosso experimento só tivemos a necessidade de truncar os sumários do RheSumaRST, e chamamos o conjunto destes sumários modificados de “RheSumaRST truncado” e os resultados da medida ROUGE-1 estão descritos na Tabela 17.

Tabela 17. Medida ROUGE para os sumários automáticos

# text id #	VeinSum	RheSumaRST	RheSumaRST truncado	GistSumm (baseline)
2000_17082	0,54545	0,53409	0,51136	0,61364
2000_17088	0,84946	0,83871	0,82796	0,38710
2000_17101	0,56436	0,75248	0,60396	0,45545
2000_17108	0,55435	0,79348	0,63043	0,56522
2000_17109	0,59551	0,50562	0,50562	0,29213
2000_17112	0,46000	0,71000	0,55000	0,22000
2000_17113	0,47328	0,57252	0,41985	0,45038
2002_22023	0,60976	0,78862	0,73984	0,25203
2003_24219	0,72477	0,66055	0,55963	0,54128
2004_26415	0,56311	0,79612	0,59223	0,29126
2005_28747	0,56044	0,56044	0,52747	0,20879
2005_28756	0,63871	0,58710	0,56129	0,41290
Média	**0,59505**	**0,66737**	**0,58073**	**0,40522**

De acordo com esses dados, a diferença na informatividade entre os sumários gerados pelo VeinSum e pelo RheSumaRST é de apenas 0,014 (na escala ROUGE), o que não representa algo muito significativo quando comparamos os dois sistemas com o baseline, que se encontra 0,18 pontos abaixo do RheSumaRST-truncado. Essa avaliação analisa a informatividade a partir de um viés estritamente objetivo, através da comparação entre sumários manuais (ideais) e automáticos. A base dessa comparação não considera os critérios subjetivos da informatividade – adotando-se, nesse sentido, a acepção sustentada por DeBeaugrande & Dressler (1981) – contemplando apenas a verificação de co-ocorrências de elementos lingüísticos (palavras) nos sumários.

9.4 Avaliação com base na informatividade

Apesar de a medida de avaliação de informatividade utilizada por Seno em seu trabalho ser reconhecidamente indicativa (medida ROUGE), apresentamos neste trabalho um critério de informatividade pautado por parâmetros notadamente subjetivos. Para tanto, partimos de uma definição clássica de informatividade, proposta por De Beaugrande & Dressler (1981),

que conceituam a mesma como um dos fatores constitutivos da natureza textual do texto. Ainda segundo os autores, a informatividade é avaliada em função das expectativas e conhecimentos dos usuários. Assim como acontece no caso de se aferir coerência e coesão ao texto, a informação de um texto não está contida, em sua totalidade, no próprio texto, sendo possível (considerado o leitor médio[61]) a depreensão de diversos fatores informativos através de processos inferenciais.

Analisando a estrutura superficial de um texto, podemos identificar, com base da própria noção de nuclearidade proposta pela RST, as unidade informativas essenciais, complementares e supérfluas, que podem ser definidas da seguinte maneira (Figura 32):

- Informação essencial: sempre deve ser incluída no sumário
- Informação complementar: pode ou não ser incluída, dependendo do modelo que o produtor tem do leitor e de sua intenção de produção. Em algumas circunstâncias, a informação complementar pode ser supérflua.
- Informação supérflua: a que pode ser descartada sem prejuízo da informação central contida no texto-fonte.

[61] O leitor médio é uma abstração correspondente a uma expectativa de produtor do texto com relação aos seus leitores. Assim, um cientista da computação, por exemplo, ao escrever um texto para sua comunidade científica, pode considerar que seu leitor médio tenha uma série de informações que não precisam ser explicitadas no texto (definições etc.).

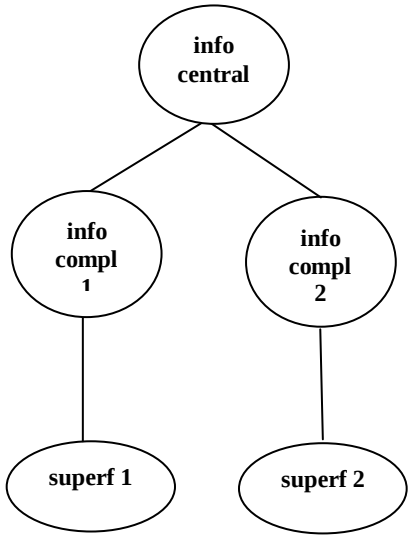
[O Instituto Nacional de Pesquisas Espaciais (Inpe) prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos, se a emissão de gases por queimadas permanecer nos níveis atuais.] ***info central*** [Esse é o pior panorama climático previsto pelo instituto,] ***info compl 1*** [disse Carlos Nobre, que participou do debate "Cenários da Amazônia", na 52ª Reunião Anual da SBPC.] ***superf 1*** [Ele afirmou que o aumento de temperatura, que pode chegar a 3C nas áreas úmidas, seria um desastre.] ***info compl 2*** [A elevação de temperatura viria acompanhada de redução das chuvas em até 15%, aumentando o risco de incêndio _inexistente há três décadas.] ***superf 2***	

Figura 32. Distribuição da informatividade

Desse modo, propomos neste trabalho um conjunto restrito de atributos aplicáveis aos sumários automáticos, em uma avaliação feita por especialistas humanos: i) texto perfeitamente informativo; ii) texto eficientemente informativo; iii) texto razoavelmente informativo; iv) texto sofrível; e v) o não-texto. Para os exemplos utilizados abaixo, foram utilizados sumários gerados automaticamente para os textos do Córpus Summ-it.

Os sumários considerados “textos perfeitamente informativos” são aqueles em que o leitor não só recupera a informação central do texto-fonte, como também as principais informações complementares que especificam o conteúdo explicitado pela informação central. Assim, um sumário informativo deve conter todas as informações relevantes de seu texto-fonte, tendo sido excluídas todas as informações menos importantes. Na Figura 33 temos um exemplo: a informação central do texto-fonte é a descoberta do dinossauro, ao passo que as informações complementares envolvem especificações sobre o animal, local de descoberta etc.

Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de uma nova espécie de dinossauro no Brasil. O animal era um carnívoro que habitou o nordeste brasileiro há 110 milhões de anos, no período Cretáceo Batizado de Santanaraptor placidus, o fóssil é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele. Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o Santanaraptor ocuparia uma posição no grupo Tyrannoraptora, o mesmo

do Tyrannossaurus rex, O exemplar de Santanaraptor encontrado pela equipe carioca foi desenterrado em 1991, mas a montagem do fóssil só foi concluída nove anos mais tarde. Tudo o que sobrou dele foram as patas e partes da cauda e da bacia, mas os pesquisadores conseguiram estimar que o bicho fosse um filhote de 1,5 metro de altura.

Figura 33. Sumário de CIENCIA_2000_17088

Os considerados "textos eficientemente informativos" correspondem aos sumários nos quais o leitor recupera a informação central do texto-fonte, mas não tem acesso a todas ou algumas informações complementares que especificam o conteúdo principal. Nesse caso, o sumário continua sendo informativo, porém priva o leitor de detalhes que, em determinados contextos, poderiam ser importantes. Dizemos que o sumário é eficientemente informativo porque a idéia central do texto fonte é recuperada e o que se omite são apenas informações de menor relevância geral. É o que temos na figura 28: a informação central (o aumento da temperatura de até 5º C nas regiões mais secas da Amazônia) está inserida, mas outras informações relevantes (o aumento de temperatura, que pode chegar a 3º C nas áreas úmidas etc.) não estão incluídas.

O Instituto Nacional de Pesquisas Espaciais prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos, Esse é o pior panorama climático previsto pelo instituto, Nobre disse que o Brasil está entre os dez países que mais poluem a atmosfera com a emissão de gás carbônico O Brasil emite 280 milhões de toneladas de carbono na atmosfera por ano. O desmatamento da Amazônia atingiu 16.926 km2 em 99, Adalberto Veríssimo, da ONG Imazon, apresentou estudo segundo o qual as cidades em regiões amazônicas ocupadas de forma predatória duram por volta de 23 anos.

Figura 34. Sumário de CIENCIA_2000_17082

Os sumários marcados como "textos razoavelmente informativos" são os que incluem apenas parte da informação central e não incluem outras informações relevantes, ou, se as incluem, não é possível relacioná-las à informação central, por esta estar incompleta. É o caso da figura 29. Neste exemplo, já temos um problema identificável na primeira sentença (A idéia foi lançada pelo ministro Roberto Amaral e detalhada ontem durante a abertura do 1º Congresso Internacional de Biodiesel): a informação central do texto (referente da

anáfora “a idéia”) não está presente. Neste caso, o leitor ainda pode inferir qual é a informação central, mas pela falta de especificação do sumário, não é possível relacionar, com segurança, as demais informações com esta proposição central.

A idéia foi lançada pelo ministro Roberto Amaral e detalhada ontem durante a abertura do 1º Congresso Internacional de Biodiesel, A intenção do governo é usar parte da soja transgênica já plantada no país, na produção do combustível. Cálculos iniciais do ministério apontam que o programa nacional do biodiesel pode representar uma economia anual de R$ 1,8 bilhão de litros de diesel importado pelo Brasile gerar 200 mil empregos no campo. que a proposta é "uma equação lógica". "Temos que ter em mente que a soja transgênica não desaparecerá no próximo ano. O ministro da Agricultura, Roberto Rodrigues, afirmou que o uso da soja transgênica "é uma boa idéia". Essa proposta,será discutida pelo governo. "Assim que tivermos uma posição,cada ministério vai tratar de sua parte", O secretário do MCT também defendeu a manutenção da produção dos transgênicos.Francelino Grando e Roberto Rodrigues chegaram ao evento de Ribeirão num microônibus movido a biodiesel. para colocar parte da frota da administração municipal movida a biodiesel, já a partir de junho.

Figura 35. Sumário de CIENCIA_2000_24219

Os sumários considerados “textos sofríveis” são aqueles nos quais a informação central simplesmente não aparece, não sendo possível ao leitor inferir qual seja a mesma. Nestes casos, o texto é apenas um conjunto de informações sem conexão lógica explicitada por qualquer elemento textual. E os “não-textos”, como o próprio termo indica, é apenas um emaranhado de fragmentos de sentenças sem qualquer conexão lógica, não sendo possível recuperar qualquer informação, mesmo que isolada do contexto. Para estes dois casos não temos exemplos no córpus.

Aspecto interessante deste modelo de avaliação é a possibilidade de podermos relacionar problemas estruturais no texto à sua informatividade, de modo que, a partir de casos genericamente considerados problemáticos (quebras de CCR, por exemplo), é possível relativizar a importância dos mesmos no contexto da ocorrência. Assim, importantes considerações acerca dos fenômenos causadores de déficits de textualidade podem ser elaboradas, explorando-se não apenas a valoração dos problemas textuais, como também apontando aspectos da própria RST referentes aos fenômenos mencionados.

Ao avaliarmos o VeinSum e o RheSumaRST utilizando estes parâmetros, aplicamos uma escala de 0 a 10 para avaliar o grau de informatividade dos sumários, considerando-se i) 0 (zero) o não-texto, ii) 2,5 o sumário sofrível, iii) 5 o sumário razoavelmente informativo, iv) 7,5 o sumário eficientemente informativo e v) 10 o sumário perfeitamente informativo. Por uma questão de contenção de esforços e de limitações de ferramental humano e recursos, os textos foram avaliados apenas por um analista humano treinado (o autor mesmo). Apesar de reconhecermos que esta limitação pode ser encarada como uma falha metodológica, tentamos minimizar o impacto da mesma usando como critério para seleção da informação central as informações destacadas nos sumários produzidos manualmente para os textos do córpus Summ-it[62]. Os resultados seguem na Tabela 18.

Tabela 18. Avaliação da qualidade textual (VeinSum e RheSumaRST)

# text id #	# texto-fonte #	VeinSum			RheSumaRST		
		extensão[1]	**CCR qbr[2]**	**informa-tividade**	**extensão**	**CCR qbr**	**informa-tividade**
2000_17082	270	100 (37,03%)	1	7,5	111 (41,11%)	3	7,5
2000_17088	370	142 (38,37%)	0	10	146 (39,45%)	1	7,5
2000_17101	315	84 (26,66%)	0	10	127 (40,31%)	1	5
2000_17108	282	90 (31,91%)	0	10	191 (67,73%)	3	5
2000_17109	241	101 (41,90%)	0	10	89 (36,92%)	1	5
2000_17112	291	93 (31,95%)	0	10	134 (46,04%)	0	7,5
2000_17113	368	111 (30,16%)	0	7,5	147 (39,94%)	0	7,5
2002_22023	377	133 (35,27%)	1	7,5	176 (46,68%)	1	5
2003_24219	360	147 (40,83%)	0	10	167 (46,38%)	4	5
2004_26415	268	80 (29,85%)	0	7,5	133 (48,50%)	2	5
2005_28747	288	95 (32,98%)	2	7,5	130 (45,13%)	0	7,5
2005_28756	455	147 (32,30%)	0	10	163 (35,82%)	4	7,5
TOTAL	**3385**	**1320**	**4**	**107,5**	**1714**	**20**	**75**
MÉDIA	**324**	**110 (34,10%)**	**0,33**	**8,95 (≈ 10)**	**143 (40,63%)**	**1,66**	**6,25 (≈ 7,5)**

(1) Número de palavras do sumário e porcentagem com relação ao tamanho do texto-fonte.
(2) Quebras de cadeias de co-referência.

[62] Sumários produzidos por especialistas em geração de resumos.

Observamos, primeiramente, que os sumários gerados pelo RheSumaRST são sensivelmente maiores que os gerados pelo VeinSum – para taxa de compressão de 30%, o primeiro apresenta sumários com uma média de 40,63% de extensão, enquanto o segundo apresenta 34,10%. Isso implica, além dos problemas referentes à violação acima do aceitável da taxa de compressão, o aumento da probabilidade de seleção de mais informação do texto-fonte (o que devemos considerar mais à frente, ao observarmos os dados referentes à avaliação automática da informatividade).

Outro aspecto avaliado no tocante à qualidade textual dos sumários é a coesão referencial, mais precisamente as quebras de cadeias de co-referência. Como vemos na tabela 16, os sumários produzidos pelo RheSumaRST apresentam 20 ocorrências de quebras, contra apenas quatro casos nos sumários produzidos pelo VeinSum. Estes resultados são interessantes porque, em tese, ambos os sistemas possuem o mesmo fundamento teórico voltado à manutenção das cadeias de co-referência, diferindo apenas no modelo de cálculo de dependência. Nossa hipótese para explicar esta diferença significativa entre o número de quebras dos sistemas é que o RheSumaRST, ao inserir mais informação, aumenta a possibilidade de inserção no sumário de elementos lingüísticos dependentes de referentes não inseridos. Essa característica do RheSumaRST, além dos problemas de referenciação, parece também ser a causadora de outros problemas estruturais nos sumários, tais como orações fragmentadas e outras anormalidades sintáticas.

Recuperando os dados da avaliação subjetiva da informatividade, assim como na utilização da ferramenta ROUGE, melhores resultados com o sistema VeinSum (média 7,5 para o VeinSum contra 5,0 para o RheSumaRST – vide tabela 16). A disparidade da extensão das diferenças entre os sistemas (VeinSum: 7,5 (subjetiva)/0,59505 (Rouge); RheSumaRTS: 5,0 (subjetiva)/0,58703 (Rouge)) reflete as diferenças na metodologia, ou seja, o nível de análise que é feito em um caso e outro, não sendo possível, portanto, compararmos os resultados dos dois métodos.

Ao passo que a medida ROUGE realiza uma avaliação superficial, comparando a co-ocorrência de termos nos sumários (automáticos e de referência), a medida subjetiva leva

em consideração diversos outros fatores preponderantes na construção da informatividade textual – dentre eles a manutenção dos elos referenciais.

No sumário do texto CIENCIA_2000_17108 (gerado pelo RheSumaRST), por exemplo, observemos a primeira sentença[63]. A expressão grifada remete exatamente ao sujeito omitido na primeira sentença (anomalia sintática) e indica uma quebra de referência, pois o referente de "essa criatura" não pode ser recuperado.

VeinSum	RheSumaRST
Um ser que invade corpos e domina a mente alheia, não é mero personagem de ficção. Para uma aranha da Costa Rica, essa criatura existe. O biólogo William Eberhard, da Universidade da Costa Rica, descobriu que as larvas desse inseto, provocam mudanças no comportamento da hospedeira. "É uma descoberta e tanto", disse o psicólogo César Ades, da USP, especialista em comportamento de aranhas. "É a primeira vez que se vê uma interação química tão complexa entre parasita e hospedeiro", A exploração alheia não tem limites. Nem mesmo no reino animal.	não é mero personagem de ficção. Para uma aranha da Costa Rica, essa criatura existe. Apesar do nome o tal invasor de corpos é só uma vespa. O biólogo William Eberhard, da Universidade da Costa Rica, descobriu que as larvas desse inseto, provocam mudanças no comportamento da hospedeira. A larva induz quimicamente a aranha a modificar o formato da própria teia Não satisfeita com a manipulação,ainda mata e devora sua anfitriã. A relação espúria começa no abdome da aranha, A larva passa de 7 a 14 dias ali dentro, Então, libera uma droga ainda desconhecida na corrente sanguínea da vítima. A substância atinge o sistema nervoso da aranha. Dopada, ela passa a repetir um único padrão de teia, Sem saber,o aracnídeo está providenciando o suporte perfeito para o casulo da parasita. Na noite em que a teia fica pronta, a larva irrompe do corpo da aranha, Para completar a exploração,ela devora sua ex-hospedeira. Só então começa a entrar no casulo, "É uma descoberta e tanto", "É a primeira vez que se vê uma interação química tão complexa entre parasita e hospedeiro", A exploração alheia não tem limites.Nem mesmo no reino animal.

Figura 36. Sumários automáticos para o texto CIENCIA_2000_17108

Outro exemplo, talvez ainda mais indicativo das diferenças entre os sistemas, é o que podemos encontrar no texto CIENCIA_2004_26415. Observemos os sumários gerados pelos dois sistemas:

CIENCIA_2004_26415	
VeinSum	**RheSumaRST**
[Para um desavisado parece até obsessão freudiana,]1 [mas	Para um desavisado parece até obsessão freudiana,

[63] Os casos de quebra encontram-se sublinhados nos sumários gerados por ambos os sistemas – vide apêndice C.

Hendrik Poynar está pedindo a todos os seus conhecidos a maior quantidade de fezes possível]2 [Bioantropólogo da Universidade MacMaster, no Canadá, está prestes a investigar a relação entre neandertais e humanos modernos]4 "Estamos recolhendo amostras de coprólitos de duas cavernas em Israel com 40 mil anos, Poynar também disse estar apostando todas as suas fichas para a melhor compreensão da evolução humana na chamada paleoproteômica _o estudo das proteínas em fósseis.	Bioantropólogo da Universidade MacMaster, no Canadá, está prestes a investigar a relação entre neandertais e humanos modernos" Estamos recolhendo amostras de coprólitos de duas cavernas em Israel com 40 mil anos, há grandes chances de elas terem preservado mais DNA do que o que se pode extrair de ossos, bem como proteínas e outras moléculas. Poynar pretende usar esse material, Os sedimentos da caverna, também vão ser peneirados."Depois disso, o que você faz é basicamente sequenciar tudo o que está ali e examinar toda a cadeia de relações alimentares, ecológicas e de parentesco das pessoas e animais que habitaram a caverna",Poynar também disse estar apostando todas as suas fichas para a melhor compreensão da evolução humana na chamada paleoproteômica_ o estudo das proteínas em fósseis.

Figura 37. Sumários automáticos para o texto CIENCIA_2004_26415

Logo no início do sumário produzido pelo RheSumaRST (Figura 37), podemos observar uma quebra de referência bastante significativa: o sistema omitiu a EDU “mas Hendrik Poynar está pedindo a todos os seus conhecidos a maior quantidade de fezes possível”, fundamental para a compreensão da expressão referencial “Bioantropólogo da Universidade MacMaster, no Canadá”, introduzida logo no início do sumário, bem como para outras menções ao pesquisador, feitas duas vezes ao longo do sumário através do sobrenome do autor (Poynar). Ao observarmos a estrutura RST do texto-fonte (Figura 38, abaixo)[64], vemos que a informação mais relevante encontra-se na EDU 4 e que a informação omitida no sumário do RheSumaRST encontra-se no satélite desta EDU, o span 1-3. Se olharmos o *acc* da EDU 4, vemos que ele contém a EDU 2, além da própria 4, e o *acc* de 2 contém 1. Desse modo, o sistema deveria, ao selecionar 4, automaticamente selecionar 1 e 2 também – no entanto, o RheSumaRST não selecionou a EDU 2, o que acarretou a perda da referência.

[64] As EDUs circuladas são os satélites da estrutura.

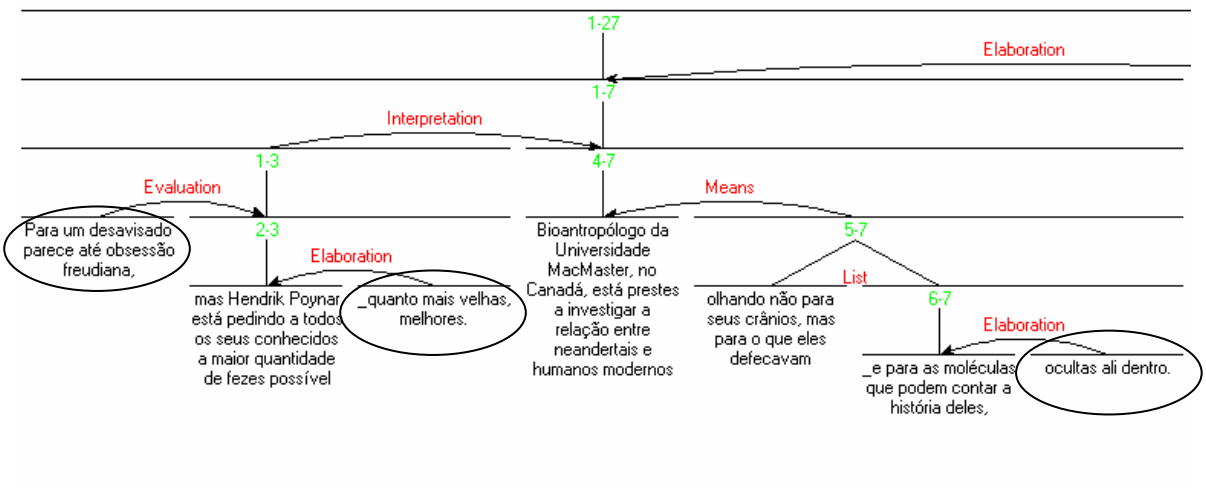


Figura 38. Subárvore do texto CIENCIA_2004_26415

Esse exemplo é interessante e oportuno, porque nos permite abordar outro caso (que guarda alguma semelhança com este) relativamente recorrente no córpus: o dos satélites que contêm uma informação relevante e cuja exclusão acarreta perda da referência (Ide & Cristea, 2000). Na Figura 38, o satélite da relação INTERPRETATION contém o termo antecedente da EDU 4; na maioria dos casos identificados no córpus, o problema encontra-se em satélites de relações ATTRIBUTION que são recuperados por expressões referenciais posteriores.

Em outros textos do córpus, verificamos ocorrências do seguinte tipo: "Esse é um alvo viável para remédios contra a obesidade", disse um dos autores, John Clapham, da empresa farmacêutica SmithKline Beecham, que fez o estudo em colaboração com a Universidade de Cambridge, Reino Unido. Estruturalmente, a citação entre aspas (o que foi dito) é núcleo com relação à fonte da citação (quem disse), satélite – uma relação do tipo ATTRIBUTION. No entanto, é muito comum que, após um nome ser introduzido como no fragmento acima, acompanhado de importantes informações credenciais (o que faz, a que instituição pertence etc.), esta mesma pessoa seja referenciada no texto apenas por seu sobrenome, ou mesmo por sua profissão – exemplo: “Clapham”, “diretor da SmithKline Beecham” etc. Em textos com este tipo de estruturação – que verificamos ser muito recorrente em textos jornalísticos e científicos – a possibilidade de o sistema de SA eliminar o satélite e causar déficits de textualidade (por perda de referência) é muito alta.

Dentre as quatro ocorrências de quebra verificadas nos sumários produzidos pelo VeinSum, três foram causadas porque os satélites de relações ATTRIBUTION não foram incluídos no *acc* de expressões referenciais das quais eram referentes. É o que temos no sumário abaixo, gerado pelo VeinSum:

O Instituto Nacional de Pesquisas Espaciais prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos, Esse é o pior panorama climático previsto pelo instituto, Nobre disse que o Brasil está entre os dez países que mais poluem a atmosfera com a emissão de gás carbônico O Brasil emite 280 milhões de toneladas de carbono na atmosfera por ano. O desmatamento da Amazônia atingiu 16.926 km2 em 99, Adalberto Veríssimo, da ONG Imazon, apresentou estudo segundo o qual as cidades em regiões amazônicas ocupadas de forma predatória duram por volta de 23 anos.

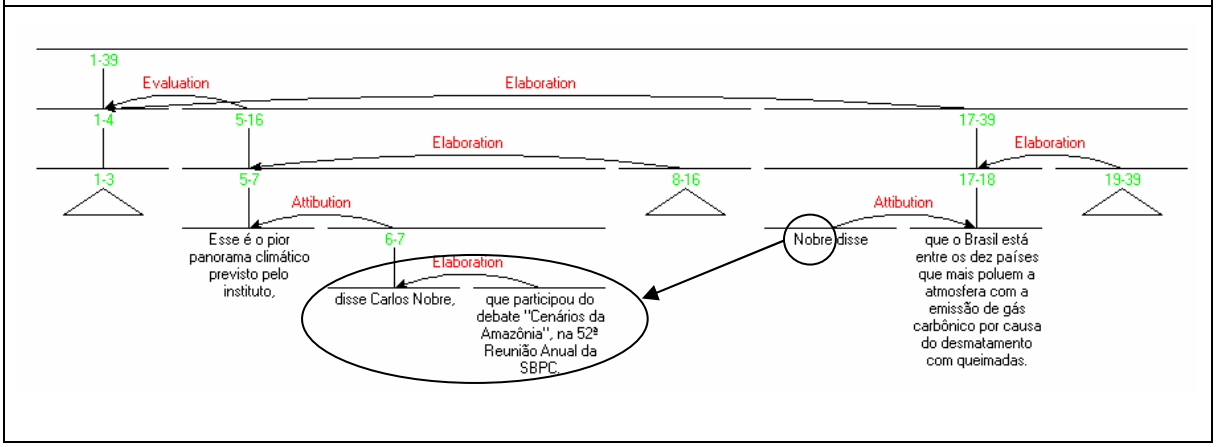


Figura 39. Sumário e estrutura RST do texto CIENCIA_2000_17082

A expressão “Nobre” recupera “Carlos Nobre”, não inserido no sumário por estar inacessível em função de sua qualidade de satélite de uma relação ATTRIBUTION (vide texto integral no apêndice B). O problema observável neste exemplo – e que foi verificado em 75% dos casos de quebra analisados entre os sumários do VeinSum – pode ser relacionado com a própria estruturação retórica. Ide & Cristea (2000) demonstram que a Teoria das Veias pode ser utilizada para indicar problemas na estruturação retórica, indicando os casos de referenciação mais complexos, tanto do ponto de vista computacional quanto do cognitivo. Segundo os autores, a existência de uma referência não compreendida no domínio de acessibilidade referencial é indicativa deste tipo de caso complexo – ou, então, de um problema estrutural.

Em exemplos como o da Figura 39, o problema parece ser a natureza da relação ATTRIBUTION. Para alguns autores (Skadhauge & Hardt, 2005), a relação ATTRIBUTION possui um padrão sintático rígido, no qual é possível identificar duas informações relacionadas por um verbo atributivo (como “dizer”, “falar”, “afirmar”, “assegurar”, “asseverar” etc.) ou por expressões indicativas (“de acordo com”, “segundo”, “na opinião de” etc.): i) quem ou que instituição profere a opinião (que nós denominamos de A QUEM SE ATRIBUI) e ii) o que é dito, afirmado etc. (denominado de O QUE SE ATRIBUI).

A literatura sobre RST e a classificação das relações (Marcu, 1997; 1999; Carlson & Marcu, 2001) consideram a relação ATTRIBUTION mononuclear. Marcu refinou a descrição da relação, distinguindo os casos de atribuição positiva e negativa, classificando esta última como ATTRIBUTION-N, mas, ainda assim, mononuclear. A “mononuclearidade” desta relação possui uma implicação lógica, portanto: a alta probabilidade de a informação sobre “quem disse” (a quem se atribui o que foi dito) não ser acessível no domínio de acessibilidade referencial de expressões referenciais anafóricas a ele relacionadas.

Ao analisarmos o uso da relação em nosso córpus, identificamos um número elevado de ocorrência desta relação – 50 ocorrências em doze textos (vide tabela 5). Em alguns casos, a relação ATTRIBUTION relaciona um emissor de uma opinião ou fala que não irá mais ser recuperado no discurso; em outros casos (a maioria deles), porém, a informação sobre o emissor será recuperada adiante. A Tabela 19 mostra a distribuição de ocorrências da relação ATTRIBUTION no córpus, acompanhada do número de vezes que uma informação presente no satélite dessas relações foi referenciado por anáforas textuais. Em muitos dos casos, estas anáforas aparecem como satélites em outras relações ATTRIBUTION (vide estrutura da Figura 39).

Tabela 19. Ocorrências de ATTRIBUTION no córpus

texto	ocorrências de ATTRIBUTION	referências ao satélite das relações ATTRIBUTION
17082	3	3

17088	1	2
17101	1	4
17108	2	1
17109	1	===
17112	2	4
17113	1	2
22023	2	===
24219	2	4
26415	1	4
28747	1	4
28756	1	4
TOTAL	**18**	**32**

Podemos depreender desses dados que, no contexto da utilização da Teoria das Veias para a recuperação de informação referencial em SA, o fato de a RST considerar mononuclear a relação atributiva tem como conseqüência a diminuição da precisão dos sistemas baseados nesse método de resolução anafórica. Obviamente, devemos ser cuidadosos ao propormos adaptações a uma teoria com o objetivo exclusivo de enquadrá-la em um determinado modelo de processamento, pois os pressupostos para sua construção não podem ser corrompidos – é um pressuposto da RST que a informação referente ao emissor de uma fala, opinião ou crítica deva ser considerada menos relevante que a mensagem proferida em si e é esta diferenciação que nos permite recuperar os papéis temáticos dentro da própria relação.

Analisando, porém, o caráter informativo da relação ATTRIBUTION, podemos observar aspectos interessantes que nos permitem discutir a relevância da inclusão dos satélites nos sumários gerados por sistemas de SA para textos jornalísticos e científicos. Para reforçarmos esta linha de raciocínio, recorremos a alguns pressupostos teóricos da Teoria da Comunicação que nos fornecem a base para discutir a relevância da informação satélite enquanto elemento contextualizador.

Segundo Floridi (2005), a informação é o conjunto estruturado de representações mentais codificadas (símbolos significantes) socialmente contextualizadas. Neste conceito clássico, o autor apresenta um aspecto importante da informação: a contextualização social. Podemos afirmar, nesse sentido, que diversos tipos de informação – principalmente em se

tratando de textos como o jornalístico e o científico – não podem prescindir do que definimos neste trabalho como **fator de legitimação da informação**.

Discursivamente, este elemento corresponde a um atributo que se liga à informação, influenciando o status da mesma. Observemos alguns exemplos:

> O presidente da Comissão Nacional de Ética em Pesquisa, William Saad Hossne, disse ontem, na 52ª Reunião Anual da Sociedade Brasileira para o Progresso da Ciência, que a comunidade científica internacional pode alterar a Declaração de Helsinque, sobre ética em pesquisas com humanos. (CIENCIA_2000_17101)

Podemos afirmar que o núcleo informativo do fragmento transcrito é "a comunidade científica internacional pode alterar a Declaração de Helsinque, sobre ética em pesquisas com humanos", o que é comprovado por sua estruturação RST, em que a informação aparece como núcleo na relação ATTRIBUTION, como podemos ver na Figura 40.

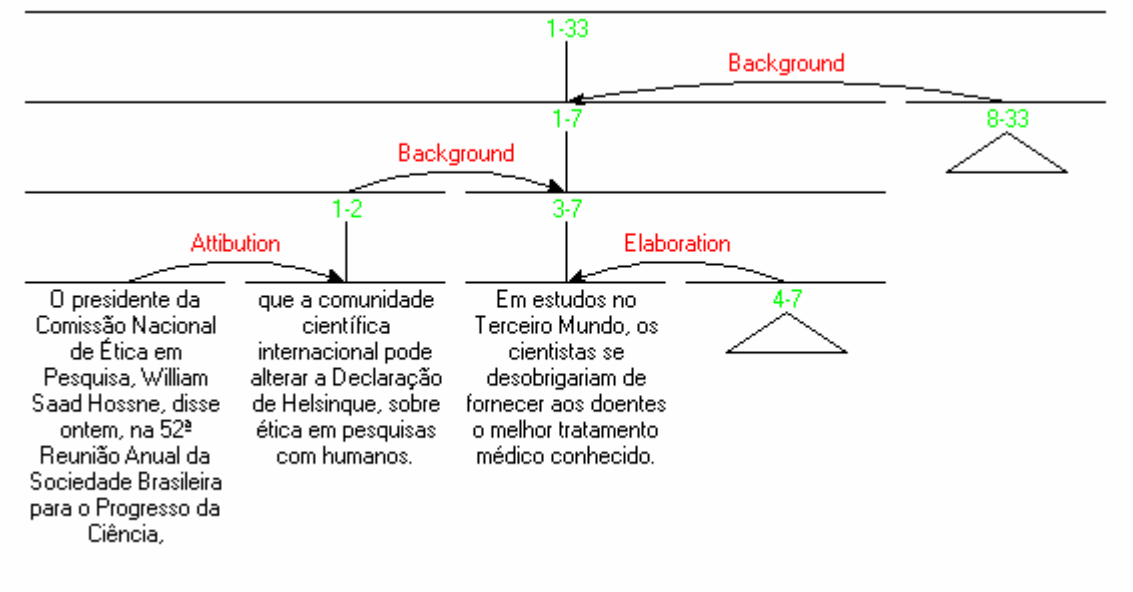


Figura 40. Estrutura RST para o texto CIENCIA_2000_17101

O satélite da mesma relação traz uma informação contextual: a quem é atribuída esta informação, onde ela foi proferida e qual a autoridade de quem a proferiu. Podemos dizer que estes dados fornecidos pelo satélite funcionam como um reforçador da validade e

credibilidade da informação, não sendo, de fato, essenciais para o conteúdo da mesma, porém importantes para que – no contexto do texto jornalístico e científico – seu objetivo comunicativo seja realizado plenamente através da aceitação por parte do público leitor. Este tipo de recurso discursivo é bastante comum e oferece ao leitor um parâmetro "de confiança" com relação à informação que lhe é transmitida. Observemos estes exemplos:

(A) "O cardiologista francês Michel de Lorgeril diz que remédios anticolesterol não são eficazes e afirma haver lobbies em favor de seu uso"[65]

(B) "Segundo o Dr. José Silva, 80, médico de uma comunidade agrícola no interior do Sergipe, os medicamentos anticolesterol são ineficazes"[66]

Nos dois exemplos acima, o objetivo comunicativo do texto é afirmar a ineficácia dos medicamentos anticolesterol, porém o fator de legitimação usado em um caso e outro diferem. Em A, temos "O cardiologista francês Michel de Lorgeril diz", o que confere certa credibilidade à informação, seja pelo fato de o médico ser um especialista, seja pela nacionalidade (que se afirma como elemento legitimador diante da crença instaurada acerca da superioridade dos países mais ricos). No exemplo B, por outro lado, o fator de legitimação é bem mais fraco, haja vista o tratar-se de um médico não-especialista muito idoso (portanto desatualizado) e por estar distante de qualquer centro de referência em pesquisa na área da medicina. Desse modo, podemos concluir que A tem muito mais possibilidades de convencer o leitor do que B, e a responsável por isso é justamente a menção ao médico, conferindo-lhe autoridade e credibilidade à informação principal. Na estrutura RST correspondente, novamente a menção aparece como satélite da relação ATTRIBUTION. Quando excluída do sumário, esse fator de legitimidade deixa de existir e, assim, a informação principal (núcleo) resta com poder comunicativo menor.

Além do problema da relevância da informação contextual e legitimadora do tópico central, temos ainda a questão da referenciação. Como podemos ver pelos dados da tabela 16, em 83,3% dos textos (10 em 12), um elemento novo no discurso é introduzido no satélite de

[65] Folha de São Paulo, 17/06/2007 (Cadermo Mais!).
[66] Exemplo fictício.

uma relação ATTRIBUTION e, posteriormente, é mencionado novamente através de algum tipo de anáfora. Esses dados, portanto, evidenciam a relevância da informação contextual contida nos referidos satélites e a necessidade de inclusão dos mesmos nos sumários automáticos.

Em vista disso, propusemos uma heurística complementar ao modelo já utilizado (ranqueamento pelo Modelo de Saliência proposto por Marcu, aliado à Teoria das Veias e a conseqüente inclusão das EDUs contidas no *acc* das EDUs selecionadas no ranqueamento), que consiste, basicamente, na inclusão do satélite da relação ATTRIBUTION no sumário toda vez que o núcleo original for incluído.

Na implementação, porém, seguimos um caminho diferente, pautados pela hipótese de que os resultados seriam os mesmos. No lugar de alterarmos o processamento de seleção das EDUs, inserimos um módulo – ao qual chamamos de Depol (despolarizador – *depolarize*) que transforma os satélites de todas as relações com a etiqueta ATTRIBUTION em núcleos, tornando a relação multinuclear apenas para a nossa aplicação – o que não implica, portanto, uma corrupção da RST.

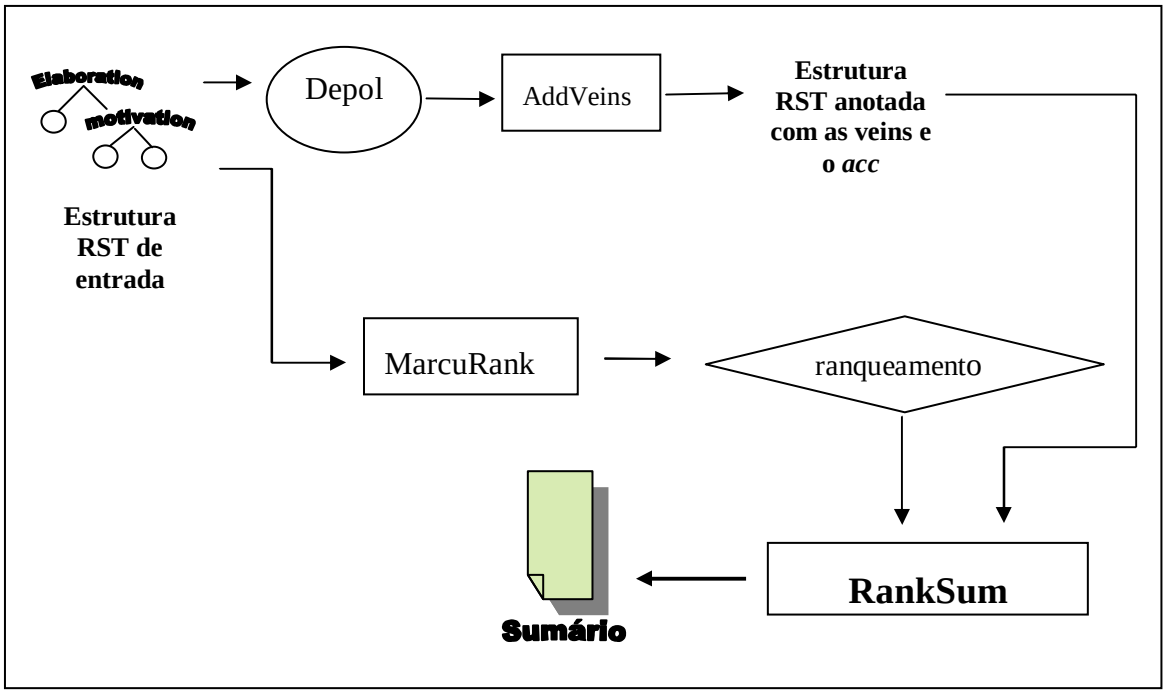


Figura 41. Arquitetura do VeinSum acrescida do despolarizador

Nessa escolha, mantemos a consistência da Teoria RST, que polariza a relação entre as duas unidades informativas que compõem a ATTRIBUTION – a informação acerca do que é dito, mencionado, pronunciado ou referido (núcleo) em oposição à fonte da mesma (satélite). De fato, na relação ATTRIBUTION, uma informação núcleo é mais relevante que o satélite, mas para os textos do gênero jornalístico analisados no córpus de estudo de nossos experimentos, verificamos que a freqüente referenciação a entidade discursiva contida nesse tipo de satélite é razão suficiente para que apliquemos a referida heurística.

Desse modo, reconhecemos como correta a afirmação da RST sobre as restrições de núcleo e satélite da relação ATTRIBUTION, mas aplicamos uma heurística que, considerada nossa aplicação – sumarização de textos com o objetivo de preservação das cadeias de co-referência – nos permite alcançar melhores resultados.

A fim de avaliar a aplicação desta heurística, os doze textos analisados anteriormente foram submetidos novamente ao VeinSum, acrescido do módulo Depol. Os resultados obtidos com relação às quebras de cadeias de co-referência podem ser observados na tabela 20 e confirmam nossa suposição inicial: os casos de quebras em elos referenciais praticamente desapareceram, limitando-se a uma única ocorrência. Entretanto, é importante considerar o fato de nosso córpus ser muito pequeno e estudos mais aprofundados em outros contextos devem ser realizados a fim de se assegurar a validade desta heurística.

tabela 20. Quebras de cadeias de co-referência nos modelos do VeinSum

textos	VeinSum	VeinSum-Depol
2000_17082	1	0
2000_17088	0	0
2000_17101	0	0
2000_17108	0	0
2000_17109	0	0
2000_17112	0	0
2000_17113	0	0
2002_22023	1	1
2003_24219	0	0
2004_26415	0	0
2005_28747	2	0
2005_28756	0	0
TOTAL	**4**	**1**

A resolução anafórica, porém, não é a única questão relevante em nossa aplicação. Como ressaltamos anteriormente, é preciso que os sumários apresentem – como evidências de sua qualidade textual – além de uma estruturação coesa e coerente, um grau satisfatório de informatividade e que respeite a taxa de compressão. Nossa avaliação, portanto, contemplou estes outros aspectos. A Figura 42 abaixo apresenta um exemplo de sumários produzidos pelos dois métodos do VeinSum:

Texto CIENCIA_2000_17082	
VeinSum	**VeinSum Depol**
O Instituto Nacional de Pesquisas Espaciais prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos, Esse é o pior panorama climático previsto pelo instituto, Nobre disse que o Brasil está entre os dez países que mais poluem a atmosfera com a emissão de gás carbônico por causa do desmatamento com queimadas. O Brasil emite 280 milhões de toneladas de carbono na atmosfera por ano. O desmatamento da Amazônia atingiu 16.926 km2 em 99, Adalberto Veríssimo, da ONG Imazon, apresentou estudo segundo o qual as cidades em regiões amazônicas ocupadas de forma predatória duram por volta de 23 anos.	O Instituto Nacional de Pesquisas Espaciais prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos, Esse é o pior panorama climático previsto pelo instituto, disse Carlos Nobre, Nobre disse que o Brasil está entre os dez países que mais poluem a atmosfera com a emissão de gás carbônico por causa do desmatamento com queimadas. O Brasil emite 280 milhões de toneladas de carbono na atmosfera por ano. O desmatamento da Amazônia atingiu 16.926 km2 em 99, Adalberto Veríssimo, da ONG Imazon, apresentou estudo segundo o qual as cidades em regiões amazônicas ocupadas de forma predatória duram por volta de 23 anos.
No. de palavras: 106 (39,25% do texto-fonte)	No. de palavras: 109 (40,36% do texto-fonte)

Figura 42. Comparação dos sumários produzidos pelos diferentes métodos do VeinSum

Como podemos observar, o sistema apenas incluiu a EDU “disse Carlos Nobre”, aumentando o sumário em três palavras e causando um pequeno corrompimento na taxa de compressão. Essa alteração, porém, além de resolver o problema da perda do referente textual de uma anáfora, não causou nenhuma modificação que prejudicasse a informatividade do sumário gerado.

Além disso, não podemos ignorar a relação que existe, principalmente no texto jornalístico – tanto o opinativo (artigos de opinião) quanto o meramente informativo (notícias em geral) – os fatos e opiniões mencionadas e sua autoria. Com base nisso, podemos considerar que a explicitação da autoria implica o aumento da informatividade.

10. Considerações finais

A presente dissertação de mestrado apresentou um amplo conjunto de estudos e experimentos realizados ao longo do desenvolvimento do projeto, que tinham por escopo central o fornecimento de subsídios lingüísticos à Sumarização Automática de textos em Língua Portuguesa, particularmente no tocante ao tratamento das cadeias de co-referência nos sumários gerados.

Nossa proposta central de pesquisa girou em torno das aplicações em conjunto da Teoria de Estruturação Retórica (RST) e da Teoria das Veias, cujo potencial já havia sido sinalizado pela implementação do RheSumaRST (Seno, 2005). Seno construiu, implementou e avaliou seu modelo sem, porém, contar com a análise e validação lingüística destas teorias (originalmente desenvolvidas para o inglês) no âmbito do português, e tampouco apresentou uma análise detalhada dos casos em que o sistema gerou sumários deficientes.

A pesquisa desenvolvida neste projeto envolveu exatamente a avaliação e aprofundamento das questões relativas às teorias envolvidas na modelagem do RheSumaRST, particularmente a Teoria das Veias. Como vimos neste trabalho, a incongruência entre a precisão apresentada pelos autores no trabalho original e aquelas verificadas nos sistemas que efetivamente a aplicam pôde ser compreendida através do cálculo de precisão que propomos neste trabalho – a precisão não-trivial – que, nos permitiu, então, lidar com predições de fato realistas para nossa modelagem de SA.

Além da referida validação de teorias, buscamos também a consecução de algumas das propostas apontadas por Seno como trabalhos futuros. Ocupamo-nos, particularmente, do acoplamento do analisador discursivo automático DiZer (Pardo, 2005) ao RheSumaRST, cujos resultados indicam como promissor o investimento nas adaptações e melhorias necessárias para a efetiva junção dos dois sistemas.

Neste trabalho, reportamos dados relevantes que orientarão o trabalho de reengenharia dos sistemas, bem como apresentamos importantes resultados de pesquisa:

- Realizamos um trabalho intenso, desde o início da pesquisa, de parceria com um cientista da computação (Jorge Marques Pelizzoni), responsável pela implementação imediata das propostas lingüísticas feitas neste trabalho;

- Procedemos a um estudo detalhado das estruturas produzidas pelo DiZer e indicamos as limitações do sistema no contexto do acoplamento ao RheSumaRST;

- Analisamos a influência do gênero textual nas aplicações de SA e elaboramos estratégias (heurísticas) de processamento para o gênero jornalístico;

- Avaliamos a Teoria das Veias e apresentamos um valor de precisão mais realista que o apresentado nos trabalhos para o inglês e outras línguas; elaboramos para isso uma medida de precisão nova na literatura: a precisão não-trivial (PNT).

- Uniformizamos, nos sistemas desenvolvidos em parceria com o referido cientista da computação, o formato de representação da informação XML, um formato que permite maior integração de nosso trabalho com o que se tem produzido na área de pesquisa (SA e correlatas);

- Desenvolvemos e utilizamos uma medida de avaliação subjetiva da informatividade;

- Descobrimos, também, que certas relações RST, como a ATTRIBUTION, têm papel fundamental na SA e, mais especificamente, na garantia do encadeamento referencial, sugerindo que satélites dessa relação devem ser considerados relevantes em sumários automáticos;

No tocante à avaliação dos sistemas individualmente (VeinSum e RheSumaRST), apresentamos também importantes observações acerca do RheSumaRST, o que nos permitiu a reimplementação da modelagem proposta por Seno em seu RheSumaRST, acrescida de especificações procedimentais novas e de uma heurística desenvolvida a partir da análise de textos jornalísticos e científicos.

Este sistema, o VeinSum, é um protótipo modificado da modelagem de Seno, e tem como entrada a estrutura RST do texto-fonte a ser sumarizado, realizando os sumários por simples justaposição das EDUs selecionadas pelo sistema. Os testes foram realizados com dados de entrada obtidos manualmente, mas os outros estudos realizados neste trabalho, somados às modificações que o DiZer tem em vista, sinalizam o potencial sucesso no acoplamento dos dois sistemas. No tocante à realização lingüística dos sumários, continuamos aguardando o desenvolvimento de modelagens de geração de língua natural que possam ser utilizadas para ter um realizador superficial real acoplado ao sistema de SA.[67]

Este trabalho possui, certamente, algumas limitações, mas, por outro lado, traz várias contribuições para a área de Sumarização Automática, apontando também diversos trabalhos futuros, conforme apresentado nas próximas seções.

10.1 Limitações

São limitações deste trabalho:

- Trabalhamos, em nossos experimentos principais, com um subconjunto de doze textos do córpus Summ-it, o que, apesar de ter proporcionado casos interessantes para a análise, não é um córpus suficientemente significativo. Optamos por um córpus mais restrito por razões de desenvolvimento do trabalho de pesquisa;

- Nossa análise de fenômenos textuais restringe-se ao gênero jornalístico, o que não nos permite estender nossas considerações a outros tipos de produção textual;

- No tocante ao estudo do fenômeno co-referencial, nos detemos apenas nos casos de expressões referenciais definidas (descrições definidas);

- O modelo de representação do conhecimento lingüístico que adotamos na manipulação dos textos, a RST, propõe uma estruturação baseada na identificação de relações entre unidades do discurso. Como a escolha é essencialmente subjetiva

[67] Atualmente, o mais promissor candidato é o trabalho apresentado por Pelizzoni (2005), ainda não concluído.

(por parte do anotador) e depende de o analista recuperar a intenção do produtor, as estruturas resultantes do processo de anotação não são definitivas, podendo ser questionadas ou alteradas, considerando-se que pode haver várias estruturas RST para um mesmo texto. Estudos mais abrangentes de nossa proposta de preservação de CCRs para outras estruturas de um mesmo texto não foram realizadas e, assim, esta proposta apresenta apenas um viés dependente da análise do córpus em questão, tanto em seu formato livre, quanto em seu formato anotado com informações retóricas;

- O algoritmo de cálculo das veias e do *acc*, utilizado para a manutenção da coesão referencial no modelo de sumarização que propomos, possui precisão de apenas 82%;

- A anotação de CCR, no estado em que a utilizamos para o processamento do córpus, ainda continha pontos controversos (entre os próprios anotadores) e algumas fragilidades. Atualmente, porém, muitos destes problemas já foram resolvidos através da uniformização das decisões dos anotadores e, para trabalhos futuros, as perspectivas são mais promissoras com relação à anotação;

10.2 Contribuições

Destacam-se, nesta seção, as principais contribuições deste trabalho, algumas delas resultando em artigos científicos, conforme citações anexas. São elas:

- Acompanhamento e análise dos dados resultantes do acoplamento dos sistemas DiZer e RheSumaRST, o que forneceu dados importantes sobre os pontos críticos a serem abordados na reengenharia dos sistemas a fim de se obter melhores resultados:
    - problema de segmentação do DiZer – ocasionado pelo uso de um tagger pouco eficiente;
  - Trabalhos decorrentes:
    - Carbonel, T. I.et al (2006).
    - Carbonel, T. I.; Rino, L. H. M. (2006a).

- Projeto e Desenvolvimento de uma ferramenta de verificação automática de quebras de Cadeias de Co-referência em sumários – o SummAlign – trabalho em conjunto com um cientista da computação.
    - Trabalho decorrente:
        - Pelizzoni, J.M. et al. (2006).

- Análise da proposta de metodologia de desenvolvimento de um sumarizador automático de estruturas RST baseado na poda de informações irrelevantes.
- Crítica do modelo de estruturação RST a partir de estudo baseado em córpus da RST, o que permitiu a elaboração de orientações de anotação relevantes aos projetos de PLN que utilizam a teoria, bem como considerações importantes acerca da relação ATTRIBUTION no gênero jornalístico
    - Trabalho decorrente:
        - Collovini, S. et al. (2007).
- Estudo e validação da Teoria das Veias para o português, o que rendeu a definição de uma medida mais realista da precisão da teoria, a precisão não-trivial.
    - Trabalho decorrente:
        - Carbonel, T. I. et al. (2007)
- Construção de córpus:
    - Córpus de sumários revisados manualmente quanto às quebras de cadeias de co-referência (resultado do acoplamento do DiZer ao RheSumaRST) – 47 sumários produzidos pelo RheSuma-2 a partir de textos jornalísticos do Córpus Rhetalho[68], com tamanho médio de 200 palavras.
        - Carbonel, T. I. et al. (2006)

[68] http://www.icmc.usp.br/~taspardo/rhetalho.html

- Córpus de textos de divulgação científica anotados retoricamente - a 10 textos da revista FAPESP, disponíveis no Córpus Lácio-WEB[69], com tamanho médio de 400 palavras.
  - Carbonel, T. I. ; Rino, L. H. M. (2006b)
- Córpus Summ-it[70] – 50 textos jornalísticos de divulgação científica do Caderno Ciência da Folha de São Paulo, com tamanho médio de 400 palavras, anotados com informação co-referencial e retórica.
  - Collovini, S.; Carbonel, T. I.; Fuchs, J. T.; Coelho, J. C.; Vieira, R.; Rino, L. H. M. (2007).

- Projeto e Desenvolvimento da reimplementação do modelo de SA de Seno e criação do protótipo VeinSum, para o qual apresentamos propostas de melhorias com relação ao sistema anterior, particularmente no tocante à manutenção da taxa de compressão, utilização da informação contextual fornecida pelo algoritmo das veias (*acc*) e manutenção dos elos co-referenciais.

10.3 Trabalhos Futuros

Como continuidade da pesquisa realizada neste projeto de mestrado, podemos sinalizar alguns desdobramentos deste trabalho:

- Replicação dos experimentos principais apresentados neste trabalho com córpus de extensão mais significativa a fim de verificarmos a consistência dos resultados obtidos para o córpus de doze textos utilizado;

- Replicação dos experimentos com a Teoria das Veias para córpus de gêneros diferentes do jornalístico, a fim de apurarmos a dependência de gênero aventada neste trabalho;

[69] http://www.nilc.icmc.usp.br/lacioweb/
[70] http://inf.unisinos.br/~renata/laboratorio/desc_corpus_Summ-it.html

- Avaliação da relação ATTRIBUTION em córpus de outros gêneros que não o jornalístico a fim de verificarmos as implicações da polarização núcleo-satélite em aplicações de SA;

- Estudo, com base em um córpus mais significativo (mais extenso e com anotação co-referencial mais abrangente), dos casos de quebra de cadeias de co-referência a fim de definir a quebra co-referencial (conceito ainda obscuro na literatura). Inicialmente, esta era uma das propostas centrais da continuidade da pesquisa (após o exame de qualificação); todavia, o trabalho com um subcórpus de apenas 12 textos do Summ-it não permitiu tal estudo, seja pelo número pequeno de quebras verificadas, seja pelo pouca diversidade de casos;

- Investigação de outras informações lingüísticas que poderiam ser agregadas à representação estrutural do discurso (RST) e ser interessantes em aplicações de SA com vistas à manutenção do elos co-referenciais. Um caminho interessante parece ser a agregação de informação semântica à anotação retórica e o estudo de padrões de relacionamento semântico entre termos referentes (antecedente e expressão referencial).

Referências Bibliográficas

Azzam, S.; Humphreys; K., Gaizauskas, R. (1999). Using coreference chains for text summarization. *ACL Workshop on Coreference and its Applications*.

Barzilay, R.; Elhadad, M. (1997). Using Lexical Chains for Text Summarization. In the *Proc. of the Intelligent Scalable Text Summarization Workshop*, Madri, Spain. Also In I. Mani and M.T. Maybury (eds.), Advances in Automatic Text Summarization.

Bentes, A C. (2001) Lingüística Textual. In: MUSSALIM, F. e BENTES, A C. (org.) *Introdução à lingüística 1*: *domínios e fronteiras*. Campinas: Cortez Editora.

Biber, D. (1988). *Variations across the speeching and writing*. Cambridge: Cambridge Press.

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

Bonini, A. (2001). Gênero textual como signo lingüístico: os reflexos da tese da arbitrariedade. In: *Linguagem em (Dis)curso*. 1(2): 123-135.

Brennan, S.; Friedman, M.; Pollard, C. (1987). A centering approach to pronouns. In *Proceedings, 25th Annual Meeting of ACL*. Stanford, pp. 155-162.

Bronckart, Jean-Paul. (1999). *Atividade de Linguagem, textos e discursos: por um interacionismo sócio-discursivo*. Tradução de Anna Rachel Machado. São Paulo: Educ.

Carbonel, T.I.; Seno, E.M.; Pardo, T.A.S.; Coelho, J.C.; Collovini, S.; Rino, L.H.M.; Vieira, R. (2006). A Two-Step Summarizer of Brazilian Portuguese Texts. Anais do IV Workshop em Tecnologia da Informação e da Linguagem Humana – TIL'2006. Ribeirão Preto – SP. Outubro.

Carbonel, T. I.; Rino, L. H. M. (2006a). RheSuma-2: análise dos sumários e estudos dos casos de quebras de cadeias de co-referência. Série de Relatório do NILC – TR 06-06, 42 pg..

Carbonel, T. I. ; Rino, L. H. M. (2006b) . Textualidade em Sumarização Automática: um estudo das cadeias de co-referência em sumários produzidos automaticamente. In: 54o. Seminário do GEL (Grupo de Estudos Lingüísticos do Estado de São Paulo), 2006, Araraquara. *Anais do 54o. Seminários do GEL, 2006*. v. 1. p. 1-1.

Carbonel, T. I.; Pelizzoni, J. M.; Rino, L. H. M. (2007). Validação preliminar da Teoria das Veias para o Português e lições aprendidas. *Anais do V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL'2007*. Franca – SP. Julho.

Carlson, L.; Marcu, D. (2001). *Discourse Tagging Reference Manual*. ISI Technical Report ISI-TR-545.

Coelho, Jorge Cesar Barbosa; Muller, Vinicius Magnus; Abreu, Sandra Collovini de; Vieira, Renata; Rino, Lucia Helena Machado (2006). Resolving Nominal Anaphora. In: *7th Workshop on the Computational Treatment of Portuguese Language, 2006*, Itatiaia. Lecture Notes in Artificial Intelligence. Berlin : Springer. v. 3960. p. 160-169.

Collovini, S.; Carbonel, T. I.; Fuchs, J. T.; Coelho, J. C.; Vieira, R.; Rino, L. H. M. (2007). Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. *Anais do V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL'2007*. Franca – SP. Julho.

Costa Val, M. G. (1991) *Redação e textualidade*. São Paulo: Martins Fontes.

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

Cristea, D. (2003). The Relationship between Discourse Structure and Referentiality in Veins Theory. In W. Menzel and C. Vertan (eds.), *Natural Language Processing between Linguistic Inquiry and System Engineering*, "Al.I. Cuza" University Publishing House, Iasi, Romênia.

Cristea, D.; Postolache, O.; Puscasu, G.; Ghetu, L. (2003). Summarizing Documents Based on Cue-phrases and References. *In the Proceedings of the International Symposium on*

Reference Resolution and its Apllications to Questions Answering and Summarization, Veneza.

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

Cunha, C; Cintra, L. F. L. (2001). *Nova Gramática do Português Contemporâneo*. Rio de Janeiro: Nova Fronteira.

De Beaugrande, R; Dressler, W. U. (1981). *Introduction to Text Linguistics*. New York: Longman.

Desiderato Antonio, Juliano. (2004). *Estrutura retórica e articulação de orações em narrativas orais e em narrativas escritas do português.* Tese de doutorado. UNESP, Araraquara.

Dias-da-Silva, B. C. (1996). *A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais.* Tese de Doutorado. UNESP, Araraquara.

Floridi, L. (2005). Semantic Conceptions of Information. In: *The Stanford Encyclopedia of Philosophy* (Edição de Inverno, 2005), Edward N. Zalta (ed.).

Grosz, B.; Joshi, A.; Weisten, S. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, V. 21, N. 2, pp. 203-225.

Halliday, M. A.K.; Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hanada, L. M.; Rino, L. H. M. (2006). RHeSumaRST: Um Software de Sumarização de Estruturas RST Baseado em Cadeias de Co-Referência. *Revista Eletrônica de Iniciação Científica*, Nro. I. Ano VI. Março, 10 p.. ISSN 1519-8219

Hirst, G. (1981). *Anaphora in natural language understanding*. Berlin: Springer Verlag.

Hoey, M. (1991). *Patterns of Lexis in Text.* Oxford: Oxford University Press.

Ide, N.; Cristea, D. (2000). A hierarchical *acc*ount of referential *acc*essibility. In: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Hong Kong, p. 416-424.

Koch, I. G. V. (1997). *Argumentação e linguagem.* Campinas: Cortez Editora.

Koch, I. G. V. (2004). *A coesão textual*. São Paulo: Contexto Editora.

Leech, G. (1983). *Principles of Pragmatics.* London: Longman.

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

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

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

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

Mann, W.C.; Matthiessen, C.;. Thompson, S.A. (1992). Rhetorical structure theory and text analysis. In: *Discourse Description: Diverse Linguistic Analyses of a Fund-raising Text*, W.C. Mann & S.A. Thompson, (editores), John Benjamins, Amsterdam/Philadelphia, pp. 39-78.

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

Marcu, D. (1999). A formal and computational synthesis of Grosz and Sidner's and Mann and Thompson's theories. In the *Proceedings of the Workshop on Levels of Representation in Discourse*, pp. 101-108. Edinburgh, Scotland.

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

Marcuschi, L. A. (1983). *Lingüistica de texto: como é e o que se faz*. Recife: Universidade Federal de Pernambuco, Série Debates 1.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communication of the Association for Computing Machinery* 38 (11), pp. 39-41.

Mitkov, R. (2002). *Anaphora Resolution*. Londres: Longman.

Morris, J.; Hirst, G. (1991). Lexical cohesion, the thesaurus, and the structure of text. *Computational Linguistics,* 17(1): 21-48.

Müller, C.; Strube, M. (2001). MMAX: A tool for the annotation of multi-modal corpora. In the *Proc. of the IJCAI 2001*, Seattle, pp. 45–50.

Müller, C., Rapp, S., and Strube, M. (2002). Applying co-training to reference resolution. In *Proc. of the 40th Annual Meeting of the ACL, Philadelphia*, PA.

Nenkova, A.; Mckeown, K. (2003) *References to Named Entities: a Corpus Study*, NAACL-HLT'03 Short Paper.

Nunes, M. G. V.; Ghiradelo, C.M.; Montilha, G.; Turine, M. (1996) Desenvolvimento de um sistema de revisão gramatical automática para o Português do Brasil. In *II Encontro para o Processamento Computacional do Português Escrito e Falado*, Curitiba.

Ng, V. and Cardie, C. (2002). Identifying anaphoric and non-anaphoric noun phrases. In *Proc. of the Nineteenth International Conference on Computational Linguistics* (COLING), Taipei, Taiwan.

O'Donnell, M. (1997). RSTTool: An RST Analysis Tool. In *Proc. of the 6th European Workshop on Natural Language Generation*, Gerhard-Mercator University, Duisburg, Alemanha.

O'Donnell, M. (2000). Rsttool 2.4: A markup tool for rhetorical structure theory. In *Proc. of the International Natural Language Generation Conference*, Mitzpe Ramon, Israel.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2002). Extractive summarization: how to identify the gist of a text. In the Proceedings of the 1st International Information Technology Symposium – I2TS, pp. 1-6. Florianópolis-SC, Brazil. October 1-5.

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

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

Pardo, T.A.S. (2005a). GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades. Série de Relatórios do NILC. NILC-TR-05-05.

Pardo, T.A.S. (2005b). Métodos para Análise Discursiva Automática. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Junho, 211p.

Pelizzoni, J. M. (2005). *Reconciliando generalidade, instanciabilidade e complexidade de Realização Lingüística por meio de Programação Concorrente por Restrições*. Monografia de qualificação em Doutorado. ICMC – USP. São Carlos-SP.

Pelizzoni, J.M.; Carbonel, T.I.; Rino, L.H.M. (2006). Constraint-Based Extract Alignment for Black-Box Evaluation of Extractive Summarization Methods. In Eric Atwell, Nancy Ide (eds.), *Proc. of the Workshop on Annotation Science: State of the Art in Enhancing Automatic Linguistic Annotation*, pp. 20-27. Held in conjunction with the *5th Conference on Language Resources and Evaluation* (LREC 2006). 24-26 MAY. Genova, Italy.

Pepineni, K.; Roukos, S.; Ward, T.; Zhu, W. (2001). BLEU: a method for automatic evaluation of machine translation**.** In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Annual Meeting of the ACL, Philadelphia, Pennsylvania.

Poesio, M., Alexandrov-Ksbadjov, M., Vieira, R., Goulart, R., and Uryupina, O. (2005). Does discourse-new detection help definite description resolution? In *Proc. of the 6th International Workshop on Computational Semantics*, Tiburg.

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

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

Seno, E.R.M. (2005). *Especificação de Heurísticas de Sumarização de Estruturas RST com Base na Preservação dos Elos Co-Referenciais.* Dissertação de Mestrado. Departamento de Computação, UFSCar.

Seno, Eloize Rossi Marques ; Rino, L. H. M. . Co-referential chaining for coherent summaries through rhetorical and linguistic modeling. In: Recent Advances in Natural Language Processing (RANLP'2005), 2005, Borovets, Bulgaria. H. Saggion (ed.), Proc. of the Workshop on Crossing Barriers in Text Summarization Research. Borovets, Bulgaria, 2005. p. 70-75.

Skadhauge, p. R.; Hardt, D. (2005). Syntactic Identication of Attribution in the RST Treebank. In: *Proceedings of Sixth International Workshop on Linguistically Interpreted Corpora (LINC-2005).* Jeju Island, Korea. p. 57-62.

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

Sparck Jones, K. (1999). Automatic Summarizing: factors and directions. In I. Mani and M. Maybury (eds.), Advances in automatic text summarization, pp. 1-12, The MIT Press.

Sporleder, C., & Lascarides, A. (2005). Exploiting linguistic cues to classify rhetorical relations. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 532-539, Borovets, Bulgaria.

Swales, J. (1992). *Genre Analysis*. Cambridge: Cambridge University Press.

Van Dijk, T.A. (1979). Recalling And Summarizing Complex Discourse. In Burghart, W. and Hölker, K., *Text Processing Textverarbeitung*. Berlin, Walter de Gruyter.

Vieira, R.; Poesio, M. (2000). An Empirically-based system for processing definite descriptions. *Computacional Linguistics*, 26(4), pp 539-593

Vieira, R.; Gorziza, F; Rossi, D.; Chishman, R.; Rossoni, R; Pinheiro, C. (2000). Extração de sintagmas nominais para o processamento de co-referência. *Anais do V Encontro para o processamento computacional da Língua Portuguesa escrita e falada – PROPOR'2000*, Atibaia-SP.

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

Vieira, R.; Gasperin, C. V.; Goulart, R. (2003). From manual to automatic annotation of coreference. In: *Proceedings International Symposium on Reference Resolution and Its Applications on Question Answering Systems*. Veneza: Universitá Ca´Foscari, v.1, p. 17-24.

ÍNDICE REMISSIVO

APÊNDICE A. Relações retóricas utilizadas neste trabalho.

Usamos neste trabalho o conjunto de 32 relações retóricas utilizado por Pardo (2005b) em seu analisador discursivo automático. DiZer. É também o elenco de relações utilizado no projeto ProCaCoSA.

Relação:	ANTITHESIS
Restrições:	
N:	o escritor julga N válido
S:	não há
N+S:	N e S estão em contraste; por causa da aparente incompatibilidade, não se pode julgar N e S válidos ao mesmo tempo; a compreensão de S e da incompatibilidade entre N e S faz o leitor aceitar melhor N
Efeito:	o leitor aceita melhor N

Figura A 1. Definição da relação ANTITHESIS

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

Figura A 2. Definição da relação ATTRIBUTION

Relação:	BACKGROUND
Restrições:	
N:	o leitor não compreenderá suficientemente N antes de ler S
S:	não há
N+S:	S aumenta a habilidade do leitor em compreender algum elemento em N
Efeito:	a habilidade do leitor para compreender N aumenta

Figura A 3. Definição da relação BACKGROUND

Relação:	CIRCUMSTANCE
Restrições:	
N:	não há
S:	apresenta uma situação (realizável)
N+S:	provê uma situação na qual o leitor pode interpretar N
Efeito:	o leitor reconhece que S provê uma situação na qual N deve ser interpretado

Figura A 4. Definição da relação CIRCUMSTANCE

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

Figura A 5. Definição da relação COMPARISON

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

Figura A 6. Definição da relação CONCESSION

Relação:	CONCLUSION
Restrições:	
N:	não há
S:	S baseia-se no que é apresentado em N
N+S:	S apresenta um fato concluído a partir da interpretação de N
Efeito:	o leitor reconhece que S é uma conclusão produzida devido à interpretação de N

Figura A 7. Definição da relação CONCLUSION

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

Figura A 8. Definição da relação CONDITION

Relação:	ELABORATION
Restrições:	
N:	não há
S:	não há
N+S:	S apresenta detalhes adicionais sobre a situação ou algum elemento de N
Efeito:	o leitor reconhece S como apresentando detalhes adicionais sobre N

Figura A 9. Definição da relação ELABORATION

Relação:	ENABLEMENT
Restrições:	
N:	apresenta uma ação do leitor não realizada
S:	não há
N+S:	a compreensão de S pelo leitor aumenta sua habilidade para realizar a ação em N
Efeito:	a habilidade do leitor para realizar a ação em N aumenta

Figura A 10. Definição da relação ENABLEMENT

Relação:	EVALUATION
Restrições:	
N:	não há
S:	não há
N+S:	S se relaciona a N pelo grau de avaliação positiva do escritor por N
Efeito:	o leitor reconhece que S avalia N e reconhece o valor que ele atribui

Figura A 11. Definição da relação EVALUATION

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

Figura A 12. Definição da relação EVIDENCE

Relação:	EXPLANATION
Restrições:	
N:	apresenta um evento ou situação
S:	não há
N+S:	S explica como e/ou porque o evento ou situação apresentado em N ocorre ou veio a ocorrer
Efeito:	o leitor reconhece que S é a razão para N ou que S explica como N ocorre

Figura A 13. Definição da relação EXPLANATION

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

Figura A 14. Definição da relação INTERPRETATION

Relação:	JUSTIFY
Restrições:	
N:	não há
S:	não há
N+S:	a compreensão de S pelo leitor aumenta sua prontidão para aceitar o direito do escritor de apresentar N
Efeito:	a prontidão do leitor para aceitar o direito do escritor de apresentar N aumenta

Figura A 15. Definição da relação JUSTIFY

Relação:	MEANS
Restrições:	
N:	uma atividade
S:	não há
N+S:	S apresenta um método ou instrumento que faz com que a realização de N seja mais provável
Efeito:	o leitor reconhece que o método ou instrumento em S faz com que a realização de N seja mais provável

Figura A 16. Definição da relação MEANS

Relação:	MOTIVATION
Restrições:	
N:	uma ação volitiva não realizada
S:	não há
N+S:	: a compreensão de S motiva a realização de N
Efeito:	o leitor reconhece que S motiva a realização de N

Figura A 17. Definição da relação MOTIVATION

Relação:	NON-VOLITIONAL CAUSE
Restrições:	
N:	apresenta uma ação não volitiva
S:	não há
N+S:	S apresenta uma situação que pode ter causado N; sem S, o leitor poderia não reconhecer o que causou a ação em N; N é mais central para a satisfação do objetivo do escritor do que S
Efeito:	o leitor reconhece a situação apresentada em S como a causa da ação apresentada em N

Figura A 18. Definição da relação NON-VOLITIONAL CAUSE

Relação:	NON-VOLITIONAL RESULT
Restrições:	
N:	não há
S:	apresenta uma ação não volitiva
N+S:	N apresenta uma situação que pode ter causado S; sem N, o leitor poderia não reconhecer o que causou a ação em S; N é mais central para a satisfação do objetivo do escritor do que S
Efeito:	o leitor reconhece a situação apresentada em N como a causa da ação apresentada em S

Figura A 19. Definição da relação NON-VOLITIONAL RESULT

Relação:	OTHERWISE
Restrições:	
N:	apresenta uma situação não realizada
S:	apresenta uma situação não realizada
N+S:	a realização de N impede a realização de S
Efeito:	o leitor reconhece que a realização de N impede a realização de S

Figura A 20. Definição da relação OTHERWISE

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

Figura A 21. Definição da relação PARENTHETICAL

Relação:	PURPOSE
Restrições:	
N:	apresenta uma ação
S:	apresenta uma situação não realizada
N+S:	S apresenta uma situação que pode realizar N
Efeito:	o leitor reconhece que a atividade em N pode ser iniciada por meio de S

Figura A 22. Definição da relação PURPOSE

Relação:	RESTATEMENT
Restrições:	
N:	não há
S:	não há
N+S:	S se relaciona a N; ambos apresentam conteúdo comparável; N é mais importante para a satisfação do objetivo do escritor
Efeito:	o leitor reconhece que S expressa o mesmo conteúdo de N, mas de forma diferente

Figura A 23. Definição da relação RESTATEMENT

Relação:	SOLUTIONHOOD
Restrições:	
N:	não há
S:	apresenta um problema
N+S:	N é uma solução para o problema em S
Efeito:	o leitor reconhece N como uma solução para o problema em S

Figura A 24. Definição da relação SOLUTIONHOOD

Relação:	SUMMARY
Restrições:	
N:	não há
S:	não há
N+S:	S apresenta o conteúdo de N resumido
Efeito:	o leitor reconhece S como um resumo do conteúdo de N

Figura A 25. Definição da relação SUMMARY

Relação:	VOLITIONAL CAUSE
Restrições:	
N:	apresenta uma ação volitiva ou uma situação que poderia surgir de uma ação volitiva
S:	não há
N+S:	S apresenta uma situação que pode ter acarretado o fato do agente da ação volitiva em N ter realizado a ação; sem S, o leitor poderia não reconhecer a motivação da ação; N é mais central para a satisfação do objetivo do escritor do que S
Efeito:	o leitor reconhece a situação apresentada em S como a causa da ação apresentada em N

Figura A 26. Definição da relação VOLITIONAL CAUSE

Relação:	VOLITIONAL RESULT
Restrições:	
N:	não há
S:	apresenta uma ação volitiva ou uma situação que poderia surgir de uma ação volitiva
N+S:	N apresenta uma situação que pode ter acarretado o fato do agente da ação volitiva em S ter realizado a ação; sem N, o leitor poderia não reconhecer a motivação da ação; N é mais central para a satisfação do objetivo do escritor do que S
Efeito:	o leitor reconhece a situação apresentada em N como a causa da ação apresentada em S

Figura A 27. Definição da relação VOLITIONAL RESULT

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

Figura A 28. Definição da relação CONTRAST

Relação:	JOINT
Restrições:	
N:	não há
Efeito:	não há

Figura A 29. Definição da relação JOINT

Relação:	LIST
Restrições:	
N:	itens comparáveis apresentados nos Ns
Efeito:	o leitor reconhece como comparáveis os itens apresentados

Figura A 30. Definição da relação LIST

Relação:	SAME-UNIT
Restrições:	
N:	os Ns apresentam informações que, juntas, constituem uma única proposição
Efeito:	o leitor reconhece que as informações apresentadas constituem uma única proposição; separadas, não fazem sentido

Figura A 31. Definição da relação SAME-UNIT

Relação:	SEQUENCE
Restrições:	
N:	as situações apresentadas nos Ns são realizadas em seqüência
Efeito:	o leitor reconhece a sucessão temporal dos eventos apresentados

Figura A 32. Definição da relação SEQUENCE

APÊNDICE B. TEXTOS DO CÓRPUS SUMM-IT ANALISADOS NOS EXPERIMENTOS REPORTADOS

CIENCIA_2000_17082

O Instituto Nacional de Pesquisas Espaciais (Inpe) prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos, se a emissão de gases por queimadas permanecer nos níveis atuais.

Esse é o pior panorama climático previsto pelo instituto, disse Carlos Nobre, que participou do debate "Cenários da Amazônia", na 52ª Reunião Anual da SBPC.

Ele afirmou que o aumento de temperatura, que pode chegar a 3C nas áreas úmidas, seria um desastre. A elevação de temperatura viria acompanhada de redução das chuvas em até 15%, aumentando o risco de incêndio _inexistente há três décadas.

Os dois fenômenos climáticos combinados levariam à desertificação de algumas áreas, disse ele.

Nobre disse que o Brasil está entre os dez países que mais poluem a atmosfera com a emissão de gás carbônico por causa do desmatamento com queimadas.

O Brasil emite 280 milhões de toneladas de carbono (sobretudo CO2, ou gás carbônico) na atmosfera por ano. Desse total, 200 milhões se devem ao desmatamento. O gás carbônico é o principal causador do efeito estufa (retenção do calor solar na atmosfera).

O desmatamento da Amazônia atingiu 16.926 km2 em 99, disse a secretária de Coordenação da Amazônia do Ministério do Meio Ambiente, Mary Allegretti. Foi melhor que em 98 (17.383 km2). "Há tendência de queda", disse.

Adalberto Veríssimo, da ONG Imazon, apresentou estudo segundo o qual as cidades em regiões amazônicas ocupadas de forma predatória duram por volta de 23 anos. Depois disso, entram em colapso total, por falta de uma política de desenvolvimento sustentável. Ele citou como exemplo as cidades de Paragominas (PA), Açailândia (MA) e Humaitá (AM).

No. de palavras: 270

CIENCIA_2000_17088

Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de uma nova espécie de dinossauro no Brasil. O animal era um carnívoro que habitou o nordeste brasileiro há 110 milhões de anos, no período Cretáceo (o último da era dos grandes répteis).

Batizado de Santanaraptor placidus, o fóssil é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele.

Para os paleontólogos, achar esse tipo de evidência equivale a acertar na loteria.

"É como se o dinossauro tivesse sido enterrado ontem", disse Alexander Kellner, geólogo do Setor de Paleovertebrados do Museu Nacional e coordenador da expedição que encontrou o fóssil na região da Chapada do Araripe, Ceará (veja mapa).

Com os tecidos preservados, os cientistas esperam poder saber mais sobre o modo de vida e a evolução dos répteis.

Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o Santanaraptor ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex, que habitou os EUA no final da era dos dinos.

Segundo Kellner, apesar de o animal ser um baixinho (poderia atingir, no máximo, 2,5 metros de altura), suas patas e bacia têm características anatômicas muito semelhantes às do ilustre réptil norte-americano.
"O Santanaraptor pode ser a espécie que deu origem ao tiranossauro 68 milhões de anos mais tarde", explicou o geólogo.
Predador
O exemplar de Santanaraptor encontrado pela equipe carioca foi desenterrado em 1991, mas a montagem do fóssil só foi concluída nove anos mais tarde. Tudo o que sobrou dele foram as patas e partes da cauda e da bacia, mas os pesquisadores conseguiram estimar que o bicho fosse um filhote de 1,5 metro de altura.
Sua estrutura óssea é de um dinossauro ágil e veloz, que provavelmente se alimentava de pequenas presas _um raptor, na linguagem dos paleontólogos. O nome é uma alusão à região onde ele viveu (a Formação Santana).

No. de palavras: 370

CIENCIA_2000_17101

O presidente da Comissão Nacional de Ética em Pesquisa, William Saad Hossne, disse ontem, na 52ª Reunião Anual da Sociedade Brasileira para o Progresso da Ciência, que a comunidade científica internacional pode alterar a Declaração de Helsinque, sobre ética em pesquisas com humanos.
Em estudos no Terceiro Mundo, os cientistas se desobrigariam de fornecer aos doentes o melhor tratamento médico conhecido. A proposta, a ser discutida, dá aos pesquisados o direito de receber terapia dada pelo governo de seu país _que pode ser nenhuma.
O debate surgiu após estudos em Ruanda e na Tailândia em que cientistas deram a grávidas com HIV um regime de AZT mais breve do que o recomendado. Queriam saber se o regime curto era melhor que nada para impedir a contaminação do feto. Para isso, outro grupo de grávidas com HIV não recebeu remédio algum. Comprovou-se que o regime mais breve basta, na maioria dos casos, para impedir a contaminação.
Os pesquisadores argumentaram que as mulheres que não receberam AZT não o teriam recebido, de qualquer forma, e que seria impossível obter resultados precisos sem esse grupo. Além disso, o resultado da pesquisa beneficia países pobres, onde o regime curto é o único acessível.
Contra esse ponto de vista, Hossne defende a norma atual: em pesquisa de tratamentos, os doentes devem receber ao menos o remédio mais eficiente já descoberto para sua doença.
Hossne citou o estudo de Tuskegee (EUA), em que negros com sífilis não foram tratados por 40 anos para que a evolução da doença fosse estudada. Os EUA, disse ele, foram um dos últimos países a assinar a Declaração de Helsinque. O texto, de 89, traça diretrizes para ética em pesquisas. Seus termos são endossados pela OMS (Organização Mundial da Saúde). A proposta já faz parte de outras declarações, como a Declaração de Consenso de Atlanta, de 99, assinadas por menos cientistas e sem endosso da OMS.

No. de palavras: 315

CIENCIA_2000_17108

Um ser que invade corpos e domina a mente alheia, forçando suas vítimas a fazer o que ele ordena, não é mero personagem de ficção. Para uma aranha da Costa Rica, essa criatura existe.
E não se trata de nenhum extraterrestre. Apesar do nome _Hymenoepimecis sp._, o tal invasor de corpos é só uma vespa.

O biólogo William Eberhard, da Universidade da Costa Rica, descobriu que as larvas desse inseto, ao parasitar a aranha Plesiometa argyra, provocam mudanças no comportamento da hospedeira.
A larva induz quimicamente a aranha a modificar o formato da própria teia para que o casulo da vespa possa se desenvolver. Não satisfeita com a manipulação, ainda mata e devora sua anfitriã.
A relação espúria começa no abdome da aranha, onde a Hymenoepimecis injeta os ovos. A larva passa de 7 a 14 dias ali dentro, fartando-se do sangue do aracnídeo, até estar madura o suficiente. Então, libera uma droga ainda desconhecida na corrente sanguínea da vítima.
A substância atinge o sistema nervoso da aranha. Dopada, ela passa a repetir um único padrão de teia, em vez de tecê-lo no formato circular tradicional. Sem saber, o aracnídeo está providenciando o suporte perfeito para o casulo da parasita.
Na noite em que a teia fica pronta, a larva irrompe do corpo da aranha, matando-a. Para completar a exploração, ela devora sua ex-hospedeira. Só então começa a entrar no casulo, onde se transformará numa vespa adulta.
"É uma descoberta e tanto", disse o psicólogo César Ades, da USP, especialista em comportamento de aranhas. "É a primeira vez que se vê uma interação química tão complexa entre parasita e hospedeiro", afirmou. A exploração alheia não tem limites. Nem mesmo no reino animal.

No. de palavras: 282

CIENCIA_2000_17109

Foi dado o primeiro passo para a diminuição das filas de espera para transplante de fígado. Cientistas britânicos detectaram, em adultos, a produção de células hepáticas a partir de células-tronco da medula óssea.
Células-tronco são células não-especializadas, capazes de dar origem a qualquer tipo de tecido. As da medula óssea dão origem a células sanguíneas. O estudo, feito por pesquisadores do Imperial College, em Londres, mostra que, além disso, elas são capazes de originar outro tipo de célula _células hepáticas_ dentro do organismo humano.
A descoberta possibilitará que pessoas com dano no fígado usem as próprias células-tronco para produzir células hepáticas. "No futuro, quando a produção de tecido hepático se tornar uma realidade, o número de transplantes poderá ser minimizado", disse à Folha por e-mail Joe Jackson, um dos autores do estudo que sai hoje na revista "Nature".
Pesquisas em camundongos haviam mostrado que células-tronco da medula óssea poderiam originar células hepáticas, além das sanguíneas. Para descobrir se o mesmo acontecia em seres humanos, os pesquisadores analisaram células do fígado de mulheres que haviam sofrido um transplante de medula óssea, cujo doador havia sido um homem.
A análise do DNA dessas células mostrou que elas continham o cromossomo Y, encontrado apenas em células masculinas. Isso indica que, de alguma forma, as células-tronco da medula óssea haviam sido capazes de "colonizar" o fígado das mulheres transplantadas. O resultado deverá contribuir para o desenvolvimento de tecidos humanos para uso terapêutico, dizem os autores.

No. de palavras: 241

CIENCIA_2000_17112

O mundo está mais seco do que se imaginava. Um estudo publicado na edição de hoje da revista "Science" afirma que 1,75 bilhão de pessoas já enfrentam severa escassez de água no planeta.

A estimativa anterior, da ONU (Organização das Nações Unidas), calculava em meio bilhão o número de indivíduos expostos atualmente ao problema.
Por severa escassez de água potável, entende-se, segundo a ONU, o uso de mais de 40% das reservas do líquido disponíveis em uma região para consumo industrial, doméstico e agrícola.
O trabalho foi coordenado pelo geocientista Charles Vörösmarty, da Universidade de New Hampshire, nos Estados Unidos.
Para realizar o cálculo, a equipe de Vörösmarty dividiu o mundo em 60 mil microrregiões. Depois, estimou a quantidade de água doce sustentável (presente em rios e reservatórios de superfície) disponível em cada região.
A projeção dos cientistas para o ano 2025 é que 3,3 bilhões de pessoas não tenham mais água para irrigação _a atividade humana que mais consome o líquido.
Todos os cálculos anteriores levavam em conta macrorregiões, como países e continentes. Eram, portanto, menos precisos.
"O que nós fizemos foi um ajuste fino", disse o pesquisador à Folha, por telefone. "Descobrimos que os recursos hídricos locais em algumas áreas estão simplesmente esgotados", afirmou.
As áreas mais atingidas, claro, são as regiões áridas do norte da África, da Ásia Central e do Oriente Médio. Mas zonas de intensa urbanização recente, como o sul dos EUA e o norte do México, também foram incluídas no novo mapa da escassez.
"A demanda aumenta de forma drástica no mundo todo", afirmou o especialista em recursos hídricos José Galizia Tundisi, do Instituto Internacional de Ecologia, em São Carlos (SP).
"As metrópoles não têm recursos hídricos suficientes para suportar o crescimento populacional", disse Tundisi.

No. de palavras: 291

CIENCIA_2000_17113

Um tratamento para a obesidade que faz você comer mais, perder peso e reduzir a taxa de gordura do corpo é o que sugere um estudo britânico publicado hoje na revista científica "Nature".
Por enquanto é só sugestão: o tratamento foi testado em camundongos. Mas os resultados levaram os cientistas a chamar o próprio estudo de "abordagem promissora" contra a obesidade.
No centro do método está um gene humano descoberto recentemente, o UCP-3, cujos mecanismos de ação ainda não são totalmente conhecidos. Sabe-se que está envolvido no processamento de energia pelas células e que um gene da mesma família, o UCP-1, está ligado à queima de gordura.
O gene UCP-3 foi inserido em camundongos e manipulado para produzir, em excesso, a proteína determinada por ele.
Os camundongos com essa alteração genética comeram até 54% mais que os camundongos normais. Apesar disso, pesavam até 23% a menos que seus companheiros. A porcentagem de tecido adiposo (gordura) sobre o volume total do corpo dos bichos também diminuiu _nos machos, em 44%; nas fêmeas, em 57%.
Sua atividade física não diferiu significativamente em relação à dos camundongos normais. Isso quer dizer que os camundongos transgênicos reduziram a gordura de seu corpo, em relação à massa muscular, sem fazer ginástica.

Os animais magros consumiram mais energia até para respirar. Em vez de armazenar a comida como gordura, transformaram-na em calor.
O fato de o UCP-3 ser um gene humano facilita aplicações clínicas da pesquisa. Um dos caminhos seria sua superestimulação.
"Esse é um alvo viável para remédios contra a obesidade", disse um dos autores, John Clapham, da empresa farmacêutica SmithKline Beecham, que fez o estudo em colaboração com a Universidade de Cambridge, Reino Unido.
Drogas baseadas no UCP-3 teriam pouco em comum com os moderadores de apetite usados hoje. "Elas funcionariam do outro lado da equação", disse Clapham. Em vez de reduzir a ingestão de energia, aumentariam seu consumo pelo corpo _o que, atualmente, é conseguido com o aumento de exercícios físicos.
Clapham não prevê fórmulas mágicas para os sedentários. Ele afirmou que dieta e exercícios devem continuar a protagonizar tratamentos para emagrecer. É preciso saber, agora, se há um limite para a superativação do gene _e quais os seus efeitos colaterais.

No. de palavras: 368

CIENCIA_2002_22023

A maioria dos cientistas concorda que os asteróides com mais de um quilômetro de diâmetro, capazes de destruir civilizações inteiras, são uma preocupação que só se justifica a cada punhado de dezenas de milhões de anos. Mas novos cálculos mostram que bólidos mais modestos, com 50 metros de diâmetro e a capacidade de destruir uma cidade, despencam do céu uma vez por milênio.
Na verdade, trata-se de boa notícia. Estimativas anteriores sugeriam que um evento desses ocorresse em média a cada 200 ou 300 anos. Os novos cálculos, aprimorados com o uso de informação antes mantida secreta pelo governo americano, oferecem uma estimativa mais precisa sobre a periodicidade desses episódios.
Durante os últimos oito anos, uma rede de satélites do Departamento de Defesa dos EUA tem monitorado a atmosfera terrestre com o objetivo de detectar explosões _obviamente na tentativa de monitorar o uso de armamento nuclear ao redor do globo.
Registros de bomba atômica nunca apareceram, mas, em compensação, o sistema foi capaz de apontar diversos eventos de explosões _todas causadas pela entrada de pequenos asteróides na atmosfera da Terra e sua subsequente quebra pelo atrito com o ar. Para os militares a coisa acabou não sendo lá muito útil, mas os dados se tornaram um prato cheio para os astrônomos.
"Em oito anos, detectamos mais de 300 eventos, graças ao nosso sistema de calibragem dos dados de satélite", conta Douglas Revelle, do Laboratório Nacional de Los Alamos, um dos autores do estudo, que está publicado na edição de hoje da revista britânica "Nature" (www.nature.com).
Incidências de rochas espaciais de poucos metros de diâmetro na atmosfera acontecem com razoável frequência _anualmente, segundo os pesquisadores. "Esses corpos medidos em metros são interessantes cientificamente, mas não oferecem absolutamente nenhum perigo aos humanos", diz Robert Jedicke, da Universidade do Arizona, escolhido pela "Nature" para comentar o estudo.
A ameaça só existe quando os bólidos têm 50 metros ou mais. Foi um meteoro desse tipo (ou um disco voador, segundo fãs de ufologia) que explodiu sobre Tunguska, na Sibéria, em 1908, destruindo centenas de quilômetros quadrados de floresta. Se um desses explodisse sobre uma região habitada, poderia matar milhões. Felizmente, com base na

nova estimativa, parece haver ainda nove séculos para catalogar os pedregulhos espaciais e se preparar para futuras colisões.

No. de palavras: 377

CIENCIA_2003_24219

Os ministérios da Agricultura e da Ciência e Tecnologia defenderam ontem o uso da soja transgênica na produção do biodiesel para abastecer parte da frota nacional de veículos.

A idéia foi lançada pelo ministro Roberto Amaral (Ciência e Tecnologia) e detalhada ontem durante a abertura do 1º Congresso Internacional de Biodiesel, realizado em Ribeirão Preto e promovido pela USP (Universidade de São Paulo) da cidade.

A intenção do governo é usar parte da soja transgênica já plantada no país, e que está com seu consumo proibido, na produção do combustível.

Cálculos iniciais do ministério apontam que o programa nacional do biodiesel pode representar uma economia anual de R$ 1,8 bilhão de litros de diesel importado pelo Brasil e gerar 200 mil empregos no campo.

Francelino Grando, secretário de Política Tecnológica Empresarial do MCT (Ministério da Ciência e Tecnologia), disse que a proposta é "uma equação lógica". "Temos que ter em mente que a soja transgênica não desaparecerá no próximo ano. E precisamos ter uma alternativa."

O ministro da Agricultura, Roberto Rodrigues, afirmou que o uso da soja transgênica "é uma boa idéia". Ele defendeu até que se continue produzindo esse tipo de soja para "esmagá-la" e transformá-la em biodiesel.

Essa proposta, ainda segundo o ministro da Agricultura, será discutida pelo governo. "Assim que tivermos uma posição, cada ministério vai tratar de sua parte", afirmou Rodrigues.

O secretário do MCT também defendeu a manutenção da produção dos transgênicos. "Um assunto produtivo não pode ser tratado pela polícia", afirmou.

Francelino Grando e Roberto Rodrigues chegaram ao evento de Ribeirão num microônibus movido a biodiesel. "É para mostrar que isso é uma realidade, que não é um sonho nem um discurso", afirmou Rodrigues.

O projeto, desenvolvido pela USP de Ribeirão, consegue produzir o biodiesel a partir da mistura de óleo vegetal _incluindo o de soja_ e etanol, álcool derivado da cana-de-açúcar.

A tecnologia difere do biodiesel utilizado em outras partes do mundo, que usa o metanol _um derivado do petróleo.

Ainda ontem, o prefeito de Ribeirão, Gilberto Maggioni (PMN), assinou uma carta de intenção com a USP para colocar parte da frota da administração municipal movida a biodiesel, já a partir de junho.

No. de palavras: 360

CIENCIA_2004_26415

Para um desavisado parece até obsessão freudiana, mas Hendrik Poynar está pedindo a todos os seus conhecidos a maior quantidade de fezes possível _quanto mais velhas, melhores. Bioantropólogo da Universidade MacMaster, no Canadá, está prestes a investigar a relação entre neandertais e humanos modernos olhando não para seus crânios, mas para o que eles defecavam _e para as moléculas que podem contar a história deles, ocultas ali dentro.

"Estamos recolhendo amostras de coprólitos [fezes fossilizadas] de duas cavernas em Israel com 40 mil anos, onde provavelmente Cro-Magnons [os primeiros humanos modernos] e neandertais viveram lado a lado", contou o pesquisador durante a reunião da AAAS (Associação Americana para o Avanço da Ciência).
Dadas as características muito especiais de preservação que as fezes podem alcançar, há grandes chances de elas terem preservado mais DNA do que o que se pode extrair de ossos, bem como proteínas e outras moléculas. Poynar pretende usar esse material, que segundo ele tende a ter aparência e consistência de chocolate, para extrair o máximo de informação possível sobre os dois grupos de humanos que habitaram a Palestina no fim da Era do Gelo.
Os sedimentos da caverna, que formam uma impressionante série temporal que vai até 150 mil anos atrás, também vão ser peneirados. "Depois disso, o que você faz é basicamente sequenciar tudo o que está ali e examinar toda a cadeia de relações alimentares, ecológicas e de parentesco das pessoas e animais que habitaram a caverna", afirma.
Poynar também disse estar apostando todas as suas fichas para a melhor compreensão da evolução humana na chamada paleoproteômica _o estudo das proteínas em fósseis.

No. de palavras: 268

CIENCIA_2005_28747

Chineses e americanos enxergam o mundo de jeitos distintos _literalmente, a julgar por uma pesquisa publicada hoje. Pesquisadores da Universidade de Michigan em Ann Arbor, nos Estados Unidos, sugerem que os olhos de asiáticos tendem a ver uma imagem no seu conjunto, prestando tanto atenção ao que está em primeiro plano quanto no fundo, enquanto os americanos demoram mais o olhar no objeto central de um quadro.
"As diferenças não são minúsculas. Depois do primeiro segundo, os americanos olharam mais para o objeto central do que para o fundo durante 600 milissegundos, enquanto isso só aconteceu por 40 milissegundos com os chineses", disse à Folha Richard Nisbett, do Departamento de Psicologia da universidade.
Ele credita à sua colega Hannah Faye Chua a idéia de testar de forma visual um dado já verificado verbalmente. Pessoas nascidas na China têm mais facilidade de se lembrar de um objeto quando o vêem pela segunda vez com o mesmo fundo que aparecia na primeira olhada _o que já não acontece com os americanos.
Se isso é verdade, em que estágio da percepção ou do processamento da imagem estaria a diferença? Foi o que o grupo testou, usando óculos que rastreiam o movimento dos olhos (veja quadro à dir.). De fato, os chineses olhavam mais para o fundo, com mais intensidade e enfocando mais áreas da imagem.
Para Nisbett, diferenças culturais _principalmente na educação das crianças_ explicariam essa assimetria. "Mães americanas tendem a usar mais substantivos, e a usar mais objetos ao brincar com seus filhos pequenos. Já as chinesas e coreanas utilizam mais verbos e enfocam mais relações sociais", afirma ele. Nisbett e Chua pretendem agora ver se diferenças como essas se manifestam entre outras culturas. O estudo está na revista "PNAS".

No. de palavras: 288

CIENCIA_2005_28756

A boa notícia é que pesquisadores argentinos acharam o mais antigo mamífero com traços modernos da América do Sul, capaz de preencher uma lacuna de milhões de anos na história desses animais. A má é que o bicho, por enquanto, não passa de um dente.

Marcelo Tejedor, paleontólogo da Universidade Nacional da Patagônia, esboça um sorriso meio sem graça quando lhe perguntam se não é uma situação frustrante. "Estamos com um saco de 200 kg de sedimento para peneirar", diz, gesticulando para indicar o tamanho do trabalho à frente. "Quem sabe não encontramos mais alguma coisa?", afirmou durante o 2º Congresso Latino-Americano de Paleontologia de Vertebrados, encerrado no fim de semana passado, no Rio de Janeiro.

A descrição do único dente, um molar inferior, foi submetida para publicação numa revista científica, mas os dados preliminares sugerem que o animal pode tanto ter sido um placentário (como humanos, cães ou baleias) quanto um marsupial (como os cangurus, que carregam seus filhotes numa bolsa). "Achamos que é mais provável que ele seja um marsupial", diz Tejedor.

Se essa hipótese for verdadeira, os pesquisadores já sabem até em que grupo marsupial encaixar o caco: ele pertenceria ao grupo dos polidolopimorfos, comedores de frutas parecidos com gambás ou cuícas que hoje estão extintos. "As cúspides [elevações] do dente indicam essa dieta", afirma Tejedor.

A importância do achado, por mais incompleto que seja, vem da sua idade. Trata-se do mais antigo mamífero sul-americano do Paleoceno, o período geológico que marca o começo do "reinado" de seu grupo no planeta, logo depois da extinção dos dinossauros, há 65 milhões de anos. No período anterior, o Cretáceo (quando os dinos ainda eram a forma dominante de vertebrado terrestre), há diversos registros de mamíferos na América do Sul, em especial na Argentina. Mas todos são formas muito primitivas, sem nenhuma relação direta com as espécies do grupo que estão vivas hoje.

A coisa muda de figura com o novo mamífero, ou o que sobrou dele. Ele foi achado em meio a sedimentos de origem marinha: pouco abaixo dele nas camadas de rocha estão mariscos fósseis que se extinguiram no fim do Cretáceo, enquanto lhe faziam companhia moluscos típicos do Paleoceno. "Isso significa que ele não é mais velho que 65 milhões de anos nem mais recente que 61,5 milhões de anos", resume o paleontólogo argentino.

A linhagem dos marsupiais e placentários (que são conhecidos pelo apelido comum de térios) se distingue justamente pelas cavidades especiais dos molares, que ajudam a triturar a comida com mais eficiência e estão presentes no espécime. Se for mesmo um marsupial, como os pesquisadores supõem, é possível que ele tenha vindo da América do Norte, onde membros do grupo aparecem bem antes no registro fóssil, durante o Cretáceo.

No. de palavras: 455

APÊNDICE C: SUMÁRIOS GERADOS PELO VEINSUM E PELO RHESUMARST

VeinSum	RheSumaRST
CIENCIA_2000_17082	
O Instituto Nacional de Pesquisas Espaciais prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos, Esse é o pior panorama climático previsto pelo instituto, Nobre disse que o Brasil está entre os dez países que mais poluem a atmosfera com a emissão de gás carbônico por causa do desmatamento com queimadas. O Brasil emite 280 milhões de toneladas de carbono na atmosfera por ano. O desmatamento da Amazônia atingiu 16.926 km2 em 99, Adalberto Veríssimo, da ONG Imazon, apresentou estudo segundo o qual as cidades em regiões amazônicas ocupadas de forma predatória duram por volta de 23 anos.	prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos. Esse é o pior panorama climático previsto pelo institut. Ele afirmou que o aumento de temperatura, seria um desastre. Nobre disse que o Brasil está entre os dez países que mais poluem a atmosfera com a emissão de gás carbônico por causa do desmatamento com queimadas. O Brasil emite 280 milhões de toneladas de carbono na atmosfera por ano. O desmatamento da Amazônia atingiu 16.926 km2 em 99, Adalberto Veríssimo, da ONG Imazon, apresentou estudo segundo o qual as cidades em regiões amazônicas ocupadas de forma predatória duram por volta de 23 anos. (AM).
No. de palavras: 106 (39,25% do texto-fonte) Avaliação: Informatividade: eficientemente informativo Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. Problema de coesão referencial – quebra de CCR. CCR: 01 quebra Nobre ➔ referente perdido por conta da relação ATTRIBUTION	No. de palavras: 111 (41,11% do texto-fonte) Avaliação: Informatividade: eficientemente informativo Textualidade: Problemas de estruturação morfossintática no texto, coesão interna ruim. Problema de coesão referencial – quebras de CCR CCR: 03 quebras Nobre ➔ referente perdido por conta da relação ATTRIBUTION
CIENCIA_2000_17088	
Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de uma nova espécie de dinossauro no Brasil. O animal era um carnívoro que habitou o nordeste brasileiro há 110 milhões de anos, no período Cretáceo Batizado de Santanaraptor placidus, o fóssil é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele. Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o Santanaraptor ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex, O exemplar de Santanaraptor encontrado pela equipe carioca foi desenterrado em 1991, mas a montagem do fóssil só foi concluída nove anos mais tarde. Tudo o que sobrou dele foram as patas e partes da cauda e da bacia, mas os pesquisadores conseguiram estimar que o bicho fosse um filhote de 1,5 metro de altura.	Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de uma nova espécie de dinossauro no Brasil.O animal era um carnívoro que habitou o nordeste brasileiro há 110 milhões de anos, no período Cretáceo Batizado de Santanaraptor placidus,o fóssil é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele.Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o Santanaraptor ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex, Segundo Kellner,apesar de o animal ser um baixinho suas patas e bacia têm características anatômicas muito semelhantes às do ilustre réptil norte-americano.mas a montagem do fóssil só foi concluída nove anos mais tarde.Tudo o que sobrou dele foram as patas e partes da cauda e da bacia,mas os pesquisadores conseguiram estimar que o bicho fosse um filhote de 1,5 metro de altura.
No. de palavras: 142 (38,37% do texto-fonte) Avaliação:	No. de palavras: 146 (39,45% do texto-fonte) Avaliação:

Informatividade: perfeitamente informativo Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. CCR: não há quebras	Informatividade: sumário eficientemente informativo. Textualidade: problema de coesão referencial – uma quebra de CCR CCR: uma quebra Kellner ➔ referente perdido por conta da relação ATTRIBUTION
CIENCIA_2000_17101	
O presidente da Comissão Nacional de Ética em Pesquisa, William Saad Hossne, disse ontem, na 52ª Reunião Anual da Sociedade Brasileira para o Progresso da Ciência, que a comunidade científica internacional pode alterar a Declaração de Helsinque, sobre ética em pesquisas com humanos. Em estudos no Terceiro Mundo, os cientistas se desobrigariam de fornecer aos doentes o melhor tratamento médico conhecido. A proposta, a ser discutida, dá aos pesquisados o direito de receber terapia dada pelo governo de seu país _que pode ser nenhuma.	que a comunidade científica internacional pode alterar a Declaração de Helsinque, sobre ética em pesquisas com humanos. Em estudos no Terceiro Mundo, os cientistas se desobrigariam de fornecer aos doentes o melhor tratamento médico conhecido. em que cientistas deram a grávidas com HIV um regime de AZT mais breve do que o recomendado. Para isso, outro grupo de grávidas com HIV não recebeu remédio algum. Os pesquisadores argumentaram que as mulheres que não receberam AZT não o teriam recebido, de qualquer forma,e que seria impossível obter resultados precisos sem esse grupo.Além disso, o resultado da pesquisa beneficia países pobres, Contra esse ponto de vista, Hossne defende a norma atual:em pesquisa de tratamentos, os doentes devem receber ao menos o remédio mais eficiente já descoberto para sua doença.
No. de palavras: 84 (26,66% do texto-fonte) Avaliação: Informatividade: perfeitamente informativo. Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. CCR: não há quebras	No. de palavras: 127 (40,31% do texto-fonte) Avaliação: Informatividade: eficientemente informativo Textualidade: Problemas de estruturação morfossintática no texto, coesão interna ruim. Problema de coesão referencial – quebras de CCR CCR: uma quebra Hossne ➔ referente perdido por conta da relação ATTRIBUTION
CIENCIA_2000_17108	
Um ser que invade corpos e domina a mente alheia, não é mero personagem de ficção. Para uma aranha da Costa Rica, essa criatura existe. O biólogo William Eberhard, da Universidade da Costa Rica, descobriu que as larvas desse inseto, provocam mudanças no comportamento da hospedeira. "É uma descoberta e tanto", disse o psicólogo César Ades, da USP, especialista em comportamento de aranhas. "É a primeira vez que se vê uma interação química tão complexa entre parasita e hospedeiro", A exploração alheia não tem limites. Nem mesmo no reino animal.	não é mero personagem de ficção. Para uma aranha da Costa Rica, essa criatura existe. Apesar do nome o tal invasor de corpos é só uma vespa. O biólogo William Eberhard, da Universidade da Costa Rica, descobriu que as larvas desse inseto, provocam mudanças no comportamento da hospedeira. A larva induz quimicamente a aranha a modificar o formato da própria teia Não satisfeita com a manipulação,ainda mata e devora sua anfitriã. A relação espúria começa no abdome da aranha, A larva passa de 7 a 14 dias ali dentro, Então, libera uma droga ainda desconhecida na corrente sanguínea da vítima. A substância atinge o sistema nervoso da aranha. Dopada, ela passa a repetir um único padrão de teia, Sem saber,o aracnídeo está providenciando o suporte perfeito para o casulo da parasita. Na noite em que a teia fica pronta, a larva irrompe do corpo da aranha, Para completar a exploração,ela devora sua ex-hospedeira. Só então começa a entrar no casulo, "É uma descoberta e tanto", "É a primeira vez que se

	vê uma interação química tão complexa entre parasita e hospedeiro", A exploração alheia não tem limites.Nem mesmo no reino animal.
No. de palavras: 90 (31,91% do texto-fonte) Avaliação: Informatividade: perfeitamente informativo. Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. CCR: não há quebras.	No. de palavras: 191 (67,73% do texto-fonte) Avaliação: Informatividade: razoavelmente informativo. Textualidade: A coesão interna foi mantida mas há problemas de coerência introduzidos por perda da coesão referencial, CCR: três quebras.
CIENCIA_2000_17109	
Foi dado o primeiro passo para a diminuição das filas de espera para transplante de fígado. Cientistas britânicos detectaram, em adultos, a produção de células hepáticas a partir de células-tronco da medula óssea. Pesquisas em camundongos haviam mostrado que células-tronco da medula óssea poderiam originar células hepáticas, além das sanguíneas. Para descobrir se o mesmo acontecia em seres humanos, os pesquisadores analisaram células do fígado de mulheres que haviam sofrido um transplante de medula óssea, A análise do DNA dessas células mostrou que elas continham o cromossomo Y, O resultado deverá contribuir para o desenvolvimento de tecidos humanos para uso terapêutico,	O estudo, mostra que, além disso, elas são capazes de originar outro tipo de célula dentro do organismo humano. Pesquisas em camundongos haviam mostrado que células-tronco da medula óssea poderiam originar células hepáticas, além das sanguíneas. Para descobrir se o mesmo acontecia em seres humanos,os pesquisadores analisaram células do fígado de mulheres que haviam sofrido um transplante de medula óssea, A análise do DNA dessas células mostrou que elas continham o cromossomo Y, O resultado deverá contribuir para o desenvolvimento de tecidos humanos para uso terapêutico, dizem os autores.
No. de palavras: 101 (41,90% do texto-fonte) Avaliação: Informatividade: perfeitamente informativo. Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. CCR: não há quebras.	No. de palavras: 89 (36,92% do texto-fonte) Avaliação: Informatividade: razoavelmente informativo. (sem a inclusão das primeiras sentenças do texto-fonte, o leitor não sabe qual é o foco da pesquisa: transplante de fígado) Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. Problemas de coesão referencial. CCR: uma quebra.
CIENCIA_2000_17112	
O mundo está mais seco do que se imaginava. Um estudo publicado na edição de hoje da revista "Science" afirma que 1,75 bilhão de pessoas já enfrentam severa escassez de água no planeta. A estimativa anterior, da ONU calculava em meio bilhão o número de indivíduos expostos atualmente ao problema. Por severa escassez de água potável, entende-se, o uso de mais de 40% das reservas do líquido disponíveis em uma região para consumo industrial, doméstico e agrícola. O trabalho foi coordenado pelo geocientista Charles Vörösmarty, da Universidade de New Hampshire, nos Estados Unidos.	O mundo está mais seco do que se imaginava. A estimativa anterior, da ONU calculava em meio bilhão o número de indivíduos expostos atualmente ao problema. Por severa escassez de água potável, entende-se,o uso de mais de 40% das reservas do líquido disponíveis em uma região para consumo industrial, doméstico e agrícola. O trabalho foi coordenado pelo geocientista Charles Vörösmarty, da Universidade de New Hampshire, nos Estados Unidos. Todos os cálculos anteriores levavam em conta macrorregiões, como países e continentes. "O que nós fizemos foi um ajuste fino", As áreas mais atingidas, claro, são as regiões áridas do norte da África, da Ásia Central e do Oriente Médio. Mas zonas de intensa urbanização recente, como o sul dos EUA e o norte do México, também foram incluídas no novo mapa da escassez. disse Tundisi.

No. de palavras: 93 (31,95% do texto-fonte) Avaliação: Informatividade: perfeitamente informativo. Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. CCR: não há quebras.	No. de palavras: 134 (46,04% do texto-fonte) Avaliação: Informatividade: eficientemente informativo. Textualidade: A coesão interna foi mantida mas há problemas de coerência introduzidos pela não seleção de informação relevante. (O sumário inclui, no início, a sentença: A estimativa anterior, da ONU calculava em meio bilhão o número de indivíduos expostos atualmente ao problema. Mas não inclui os dados atuais, o que permite ao leitor a comparação, elemento necessário para se compreender a progressão do problema, que é exatamente o foco do texto) CCR: não há quebras
CIENCIA_2000_17113	
Um tratamento para a obesidade que faz você comer mais, perder peso e reduzir a taxa de gordura do corpo é o que sugere um estudo britânico publicado hoje na revista científica "Nature". Por enquanto é só sugestão: o tratamento foi testado em camundongos. Mas os resultados levaram os cientistas a chamar o próprio estudo de "abordagem promissora" contra a obesidade. No centro do método está um gene humano descoberto recentemente, o UCP-3, cujos mecanismos de ação ainda não são totalmente conhecidos. O gene UCP-3 foi inserido em camundongos e manipulado O fato de o UCP-3 ser um gene humano facilita aplicações clínicas da pesquisa. Um dos caminhos seria sua superestimulação.	Um tratamento para a obesidade que faz você comer mais, perder peso e reduzir a taxa de gordura do corpo é o que sugere um estudo britânico publicado hoje na revista científica "Nature". Por enquanto é só sugestão: Mas os resultados levaram os cientistas a chamar o próprio estudo de "abordagem promissora" contra a obesidade. No centro do método está um gene humano descoberto recentemente, o UCP-3, O gene UCP-3 foi inserido em camundongos e manipulado Os camundongos com essa alteração genética comeram até 54% mais que os camundongos normais. Apesar disso, pesavam até 23% a menos que seus companheiros. A porcentagem de tecido adiposo sobre o volume total do corpo dos bichos também diminuiu Sua atividade física não diferiu significativamente em relação à dos camundongos normais. O fato de o UCP-3 ser um gene humano facilita aplicações clínicas da pesquisa. Um dos caminhos seria sua superestimulação.
No. de palavras: 111 (30,16% do texto-fonte) Avaliação: Informatividade: eficientemente informativo. Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. CCR: não há quebras	No. de palavras: 147 (39,94% do texto-fonte) Avaliação: Informatividade: eficientemente informativo. Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. CCR: não há quebras
CIENCIA_2002_22023	
A maioria dos cientistas concorda que os asteróides com mais de um quilômetro de diâmetro, capazes de destruir civilizações inteiras, são uma preocupação que só se justifica a cada punhado de dezenas de milhões de anos. Mas novos cálculos mostram que bólidos mais modestos, com 50 metros de diâmetro e a capacidade de destruir uma cidade, despencam do céu uma vez por milênio. Durante os últimos oito anos, uma rede de satélites do Departamento de Defesa dos EUA tem monitorado a atmosfera terrestre Registros de bomba atômica nunca	A maioria dos cientistas concorda que os asteróides com mais de um quilômetro de diâmetro, capazes de destruir civilizações inteiras, são uma preocupação que só se justifica a cada punhado de dezenas de milhões de anos. Mas novos cálculos mostram que bólidos mais modestos, com 50 metros de diâmetro e a capacidade de destruir uma cidade, despencam do céu uma vez por milênio.Na verdade, trata-se de boa notícia. Estimativas anteriores sugeriam que um evento desses ocorresse em média a cada 200 ou 300 anos. Os novos cálculos, oferecem uma estimativa

apareceram, mas, em compensação, o sistema foi capaz de apontar diversos eventos de explosões Incidências de rochas espaciais de poucos metros de diâmetro na atmosfera acontecem com razoável frequência "Esses corpos medidos em metros são interessantes cientificamente, mas não oferecem absolutamente nenhum perigo aos humanos"	mais precisa sobre a periodicidade desses episódios. Durante os últimos oito anos, uma rede de satélites do Departamento de Defesa dos EUA tem monitorado a atmosfera terrestre Registros de bomba atômica nunca apareceram, mas, em compensação, o sistema foi capaz de apontar diversos eventos de explosões Incidências de rochas espaciais de poucos metros de diâmetro na atmosfera acontecem com razoável frequência "Esses corpos medidos em metros são interessantes cientificamente,mas não oferecem absolutamente nenhum perigo aos humanos", A ameaça só existe e se preparar para futuras colisões.
No. de palavras: 133 (35,27% do texto-fonte) Avaliação: Informatividade: eficientemente informativo. Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. CCR: uma quebra	No. de palavras: 176 (46,68% do texto-fonte) Avaliação: Informatividade: razoavelmente informativo. Textualidade: Problemas de estruturação morfossintática (vida a última sentença do sumário) CCR: uma quebra
CIENCIA_2003_24219	
Os ministérios da Agricultura e da Ciência e Tecnologia defenderam ontem o uso da soja transgênica na produção do biodiesel Cálculos iniciais do ministério apontam que o programa nacional do biodiesel pode representar uma economia anual de R$ 1,8 bilhão de litros de diesel importado pelo Brasil e gerar 200 mil empregos no campo. Francelino Grando, secretário de Política Tecnológica Empresarial do MCT disse que a proposta é "uma equação lógica". "Temos que ter em mente que a soja transgênica não desaparecerá no próximo ano. O ministro da Agricultura, Roberto Rodrigues, afirmou que o uso da soja transgênica "é uma boa idéia". Essa proposta, será discutida pelo governo. "Assim que tivermos uma posição, cada ministério vai tratar de sua parte", O secretário do MCT também defendeu a manutenção da produção dos transgênicos. Francelino Grando e Roberto Rodrigues chegaram ao evento de Ribeirão num microônibus movido a biodiesel.	A idéia foi lançada pelo ministro Roberto Amaral e detalhada ontem durante a abertura do 1º Congresso Internacional de Biodiesel, A intenção do governo é usar parte da soja transgênica já plantada no país, na produção do combustível. Cálculos iniciais do ministério apontam que o programa nacional do biodiesel pode representar uma economia anual de R$ 1,8 bilhão de litros de diesel importado pelo Brasile gerar 200 mil empregos no campo. que a proposta é "uma equação lógica". "Temos que ter em mente que a soja transgênica não desaparecerá no próximo ano. O ministro da Agricultura, Roberto Rodrigues, afirmou que o uso da soja transgênica "é uma boa idéia". Essa proposta,será discutida pelo governo. "Assim que tivermos uma posição,cada ministério vai tratar de sua parte", O secretário do MCT também defendeu a manutenção da produção dos transgênicos.Francelino Grando e Roberto Rodrigues chegaram ao evento de Ribeirão num microônibus movido a biodiesel. para colocar parte da frota da administração municipal movida a biodiesel, já a partir de junho.
No. de palavras: 147 (40,83% do texto-fonte) Avaliação: Informatividade: perfeitamente informativo Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. CCR: não há quebras	No. de palavras: 167 (46,38% do texto-fonte) Avaliação: Informatividade: razoavelmente informativo. Textualidade: A coesão interna foi mantida, mas há problemas de coerência introduzidos por perda da coesão referencial, CCR: quatro quebras.
CIENCIA_2004_26415	
Para um desavisado parece até obsessão freudiana, mas Hendrik Poynar está pedindo a todos os seus	Para um desavisado parece até obsessão freudiana, Bioantropólogo da Universidade MacMaster, no

conhecidos a maior quantidade de fezes possível Bioantropólogo da Universidade MacMaster, no Canadá, está prestes a investigar a relação entre neandertais e humanos modernos "Estamos recolhendo amostras de coprólitos de duas cavernas em Israel com 40 mil anos, Poynar também disse estar apostando todas as suas fichas para a melhor compreensão da evolução humana na chamada paleoproteômica _o estudo das proteínas em fósseis.	Canadá, está prestes a investigar a relação entre neandertais e humanos modernos" Estamos recolhendo amostras de coprólitos de duas cavernas em Israel com 40 mil anos, há grandes chances de elas terem preservado mais DNA do que o que se pode extrair de ossos, bem como proteínas e outras moléculas. Poynar pretende usar esse material, Os sedimentos da caverna, também vão ser peneirados."Depois disso, o que você faz é basicamente sequenciar tudo o que está ali e examinar toda a cadeia de relações alimentares, ecológicas e de parentesco das pessoas e animais que habitaram a caverna",Poynar também disse estar apostando todas as suas fichas para a melhor compreensão da evolução humana na chamada paleoproteômica_ o estudo das proteínas em fósseis.
No. de palavras: 80 (29,85% do texto-fonte) Avaliação: Informatividade: eficientemente informativo Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. CCR: não há quebras	No. de palavras: 133 (48,50% do texto fonte) Avaliação: Informatividade: eficientemente informativo Textualidade: Problemas de estruturação morfossintática das sentenças e perda da coesão referencial. CCR: duas quebras
CIENCIA_2005_28747	
Chineses e americanos enxergam o mundo de jeitos distintos Pesquisadores da Universidade de Michigan em Ann Arbor, nos Estados Unidos, sugerem que os olhos de asiáticos tendem a ver uma imagem no seu conjunto enquanto os americanos demoram mais o olhar no objeto central de um quadro. Ele credita à sua colega Hannah Faye Chua a idéia de testar de forma visual um dado já verificado verbalmente. Para Nisbett, diferenças culturais explicariam essa assimetria. Nisbett e Chua pretendem agora ver se diferenças como essas se manifestam entre outras culturas. O estudo está na revista "PNAS".	Chineses e americanos enxergam o mundo de jeitos distintos Pesquisadores da Universidade de Michigan em Ann Arbor, nos Estados Unidos, sugerem que os olhos de asiáticos tendem a ver uma imagem no seu conjunto, enquanto os americanos demoram mais o olhar no objeto central de um quadro."As diferenças não são minúsculas. Depois do primeiro segundo, os americanos olharam mais para o objeto central do que para o fundo durante 600 milissegundos, enquanto isso só aconteceu por 40 milissegundos com os chineses",disse à Folha Richard Nisbett, do Departamento de Psicologia da universidade. Ele credita à sua colega Hannah Faye Chua a idéia de testar de forma visual um dado já verificado verbalmente. em que estágio da percepção ou do processamento da imagem estaria a diferença? Foi o que o grupo testou,
No. de palavras: 95 (32,98% do texto fonte) Avaliação: Informatividade: eficientemente informativo Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático, mas há problemas de coesão referencial, CCR: duas quebras. Nisbett ➔ causado pela relação ATTRIBUTION	No. de palavras: 130 (45,13% do texto-fonte) Avaliação: Informatividade: eficientemente informativo Textualidade: Problemas de estruturação morfossintática das sentenças. CCR: não há quebras.
CIENCIA_2005_28756	
A boa notícia é que pesquisadores argentinos acharam o mais antigo mamífero com traços modernos da América do Sul, capaz de preencher	"Quem sabe não encontramos mais alguma coisa?", A descrição do único dente, um molar inferior, foi submetida para publicação numa revista científica,

uma lacuna de milhões de anos na história desses animais. A má é que o bicho, por enquanto, não passa de um dente. A descrição do único dente, um molar inferior, foi submetida para publicação numa revista científica, mas os dados preliminares sugerem que o animal pode tanto ter sido um placentário quanto um marsupial A importância do achado, vem da sua idade. Trata-se do mais antigo mamífero sul-americano do Paleoceno, o período geológico que marca o começo do "reinado" de seu grupo no planeta, logo depois da extinção dos dinossauros, há 65 milhões de anos. A linhagem dos marsupiais e placentários se distingue justamente pelas cavidades especiais dos molares, Se for mesmo um marsupial, é possível que ele tenha vindo da América do Norte,	mas os dados preliminares sugerem que o animal pode tanto ter sido um placentário quanto um marsupial A importância do achado, vem da sua idade.Trata-se do mais antigo mamífero sul-americano do Paleoceno, o período geológico que marca o começo do "reinado" de seu grupo no planeta, logo depois da extinção dos dinossauros, há 65 milhões de anos. Mas todos são formas muito primitivas, sem nenhuma relação direta com as espécies do grupo que estão vivas hoje. A coisa muda de figura com o novo mamífero, ou o que sobrou dele. Ele foi achado em meio a sedimentos de origem marinha: A linhagem dos marsupiais e placentários se distingue justamente pelas cavidades especiais dos molares, Se for mesmo um marsupial,é possível que ele tenha vindo da América do Norte, onde membros do grupo aparecem bem antes no registro fóssil, durante o Cretáceo.
No. de palavras: 147 (32,30% do texto-fonte) Avaliação: Informatividade: perfeitamente informativo Textualidade: As sentenças selecionadas estão íntegras do ponto de vista morfossintático e a coesão interna foi mantida. CCR: não há quebras	No. de palavras: 163 (35,82% do texto-fonte) Avaliação: Informatividade: eficientemente informativo. Textualidade: A coesão interna foi mantida mas há problemas de coerência introduzidos pela não seleção de informação relevante. Problemas de coesão referencial. CCR: quatro quebras.