UNIVERSIDADE FEDERAL DE SÃO CARLOS

DEPARTAMENTO DE LETRAS

PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

O USO DE INFORMAÇÕES SEMÂNTICAS DO PALAVRAS: EM BUSCA DO APRIMORAMENTO DA SELEÇÃO DE UNIDADES TEXTUAIS CORREFERENTES NA SUMARIZAÇÃO AUTOMÁTICA

Élen Cátia Tomazela

Orientadora: Dra. Lucia Helena Machado Rino

São Carlos

Junho/2010

UNIVERSIDADE FEDERAL DE SÃO CARLOS

DEPARTAMENTO DE LETRAS

PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

O USO DE INFORMAÇÕES SEMÂNTICAS DO PALAVRAS: EM BUSCA DO APRIMORAMENTO DA SELEÇÃO DE UNIDADES TEXTUAIS CORREFERENTES NA SUMARIZAÇÃO AUTOMÁTICA

Élen Cátia Tomazela

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Linguística da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Linguística.

São Carlos

Junho/2010

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

T655ui

Tomazela, Élen Cátia.
O uso de informações semânticas do PALAVRAS : em busca do aprimoramento da seleção de unidades textuais correferentes na Sumarização Automática / Élen Cátia Tomazela. -- São Carlos : UFSCar, 2010.
135 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2010.

1. Linguística - processamento de dados. 2. Sumarização automática. 3. Textualidade. 4. Correferência. I. Título.

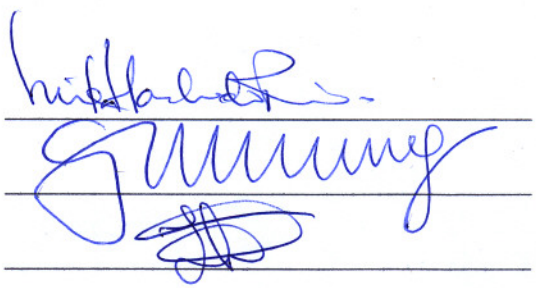
CDD: 410.285 (20ª)

BANCA EXAMINADORA

Profa. Dra. Lucia Helena Machado Rino

Profa. Dra. Maria Eduarda Giering

Prof. Dr. Thiago Alexandre Salgueiro Pardo



AGRADECIMENTOS

A Deus, por me devolver a saúde para que eu pudesse desenvolver este trabalho e pelas lições ensinadas à força, mesmo que eu lutasse muitas vezes contra a Sua vontade.

A minha mãe, que acompanhou a minha caminhada, passo a passo, até essa conquista.

A minha orientadora, Profa. Lucia Rino, por sua dedicação e direcionamentos que tornaram possível a realização deste trabalho.

A Prof. Eduarda Giering, ao Prof. Bento Dias da Silva e ao Prof. Thiago Pardo, pelas contribuições na qualificação e na defesa dessa dissertação. Especialmente o Prof. Thiago, pelas constantes contribuições ao desenvolvimento deste trabalho.

A todos os meus amigos do mestrado e doutorado do Departamento de Computação da UFSCar, que me acolheram como sendo uma deles durante todo o período em que eu estive "em departamento alheio".

A minha amiga e "irmã" Viviane Mazo, que todo santo ano, desde o início do meu mestrado, cuidou das minhas "férias", sempre me arrumando alguma viagem prazerosa prá fazer para que eu pudesse recuperar as energias e seguir adiante. Além também da sua amizade constante desde 1998, segurando sempre a minha barra.

A minha amiga Patrícia Nunes Gonçalves, que mesmo do outro lado do Atlântico, acompanhou tão de perto o processo, como se estivesse ali, na porta ao lado.

Aos meus colegas do LaLiC – Laboratório de Linguística Computacional, especialmente ao Josué, pelos momentos de descontração. Também aos colegas do NILC – Núcleo Interinstitucional de Linguística Computacional, pela amizade e contribuições ao longo da pesquisa.

A Claudia Dias de Barros e a Daniel Saraiva Leite, pela ajuda com as atividades realizadas.

Ainda, agradecimentos especiais ao Flávio, ao Danilo e à Mayra pelos cafés e comilanças ao redor da cidade, jantares em casa, jogos de boliche, etc...

A FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo, pelo suporte financeiro a este trabalho.

RESUMO

Esta dissertação tem como foco a proposta de um modelo heurístico teórico que utiliza, além da Teoria das Veias, informações semânticas provenientes do *Parser* PALAVRAS para melhorar a seleção de unidades correferentes para a inclusão em sumários automáticos. A partir da análise dos problemas apresentados pelo sumarizador automático VeinSum, identificou-se a necessidade de melhorar a saliência dos sumários produzidos, além de reduzir o tamanho dos mesmos para que se aproximassem mais da taxa de compressão ideal. Propõe-se, então, a eliminação de unidades textuais de importância secundária no que tange à clareza referencial, sem danificá-la, no entanto. Para isso, heurísticas baseadas nas informações semânticas do PALAVRAS foram propostas. Apesar de o *parser* apresentar inconsistências de etiquetação semântica, a anotação de todos os sintagmas nominais dos 50 textos-fonte que compõem o *corpus* Summ-it foi pós-editada manualmente para melhorar a confiabilidade das heurísticas geradas. Foram analisados 11 textos pertencentes ao *corpus* e os resultados são satisfatórios, porém reconhece-se que, para melhor avaliar os resultados desta proposta, faz-se necessário um estudo mais amplo.

ABSTRACT

This dissertation aims at presenting a theoretical heuristic model which not only takes into consideration the Veins Theory, but also semantic information obtained from the *Parser* PALAVRAS to improve the selection of correferential textual units to be included in automatic summaries. Based on the analysis of the problems presented by VeinSum, an automatic summarizer, two main issues have been raised: the necessity of improving its summaries salience and reducing their size so that they suit the compression rate more adequately. Better results can be achieved through the elimination of irrelevant textual units although the summaries referential clarity may not be damaged. Heuristics based on the semantic information have then been proposed. Despite the semantic annotation inconsistencies, all the noun phrases that compose the Summ-it *Corpus* have been post-edited manually, which increases the credibility of the heuristics. Eleven texts from the *corpus* have been analysed and the results obtained are satisfactory, although a wider study would be required to better evaluate the results of this proposal.

ÍNDICE GERAL[1]

1 INTRODUÇÃO 1

2 A SUMARIZAÇÃO AUTOMÁTICA 6

3 CONCEITOS LINGUÍSTICOS RELEVANTES PARA ESTA PESQUISA 9

3.1 Textualidade 9

3.1.1 Coerência Textual 10

3.1.2 Coesão Textual 11

3.2 Cadeias de Correferência 13

4 PRESSUPOSTOS TEÓRICOS DO PROJETO 16

4.1 *Rhetorical Structure Theory* (RST) – Teoria de Estruturação Retórica 16

4.2 *Veins Theory* (VT) - Teoria das Veias 20

4.3 Modelo de Saliência 22

5 O VEINSUM – DETALHAMENTO E CRÍTICA DO SISTEMA 25

5.1 Tipos de problemas apresentados pelos sumários do VeinSum 28

5.1.1 Ausência da EDU com antecedente mais completo 28

5.1.2 Rebaixamento da classificação da saliência em prol da taxa de compressão 30

5.1.3 Corrompimento da taxa de compressão 31

5.2 Inconsistências entre os pressupostos do sistema e os sumários gerados 32

5.2.1 Incoerências entre os próprios resultados intermediários 33

5.2.2 Incoerências entre os dados intermediários do sistema e o sumário 33

5.2.3 Casos de exclusão da Relação de ATTRIBUTION 36

6 O USO DE INFORMAÇÕES SEMÂNTICAS DO PALAVRAS: EM BUSCA DO APRIMORAMENTO DA SELEÇÃO DE UNIDADES CORREFERENTES NA SUMARIZAÇÃO AUTOMÁTICA 40

6.1 *Parser* PALAVRAS 40

[1] Utiliza-se a palavra índice nesta página, pois sumário, neste trabalho, significa resumo.

6.2 Dados sobre a pós-edição manual da etiquetagem semântica do PALAVRAS para o *Corpus* Summ-it 48
6.2.1 Decisões de Pós-edição 48
6.2.2 Problemas de segmentação 50
6.2.3 Problemas de etiquetação 51
6.2.4 Problemas de desambiguação 53
6.2.5 Crítica de desempenho do PALAVRAS 54
6.2.6 Síntese numérica dos casos 57
6.3 Heurísticas de seleção de EDUs 59
6.4 Detalhamento do modelo proposto 64
6.5 Algoritmo de aprimoramento de seleção de unidades correferentes na SA 69
6.6 Aplicação do algoritmo, geração e análise dos novos sumários 71
6.6.1 Sumários novos que não tiveram a saliência rebaixada 72
6.6.2 Sumários novos que apresentam TC mais próxima da ideal 81
6.6.3 Substituição de EDU proveniente de *acc* por EDU altamente saliente 92
6.6.4 Sumário novo que comprova a necessidade de rebaixamento de saliência 95
6.6.5 Impossibilidade de aplicação do modelo heurístico 96
6.6.6 Síntese da aplicação das heurísticas para a seleção de unidades textuais correferentes 97
7 CONSIDERAÇÕES FINAIS 102
REFERÊNCIAS BIBLIOGRÁFICAS 104
APÊNDICE A. TEXTOS-FONTE DO CÓRPUS SUMM-IT ANALISADOS 109
APÊNDICE B. ELENCO DE PROTÓTIPOS SEMÂNTICOS PROVIDOS PELO PALAVRAS 116
APÊNDICE C. CONJUNTO DE REGRAS DE ASSOCIAÇÃO 127
APÊNDICE D. CONJUNTO DE REGRAS DE CLASSIFICAÇÃO 131
APÊNDICE E. EXEMPLOS DE ANOTAÇÃO RST INCONSISTENTE 132

ÍNDICE DE FIGURAS

Figura 1. Arquitetura geral de um sistema de SA de abordagem profunda.....7

Figura 2. Excerto e estrutura RST que evidenciam uma relação mononuclear.....16

Figura 3. Excerto e estrutura RST que evidenciam uma relação multinuclear 16

Figura 4. Definição da relação PURPOSE 17

Figura 5. Definição da relação SEQUENCE.....17

Figura 6. Excerto do texto CIENCIA_2000_6391 19

Figura 7. Árvore RST do excerto da Figura 6 19

Figura 8. Árvore RST anterior anotada com o *acc* de cada EDU 21

Figura 9. Estrutura RST do texto CIENCIA_2001_6423 23

Figura 10. Arquitetura do VeinSum 25

Figura 11. Estrutura RST que compreende N e S de uma relação de ATTRIBUTION.....26

Figura 12. Estrutura RST que ilustra N e S da relação retórica ATTRIBUTION.....37

Figura 13. Estrutura RST correspondente a N e S do caso acima 38

Figura 14. Exemplo de anotação completa fornecida pelo PALAVRAS.....41

Figura 15. Estrutura sintática arbórea de parte do excerto da Figura 6 42

Figura 16. Árvore binária de decisão com os traços semânticos.....43

Figura 17. Protótipos relacionados a plantas 44

Figura 18. Protótipos referentes a itens lexicais semanticamente muito distintos 46

Figura 19. Arquitetura do modelo proposto 65

Figura 20. Ilustração da mensagem completa dada pela relação CIRCUMSTANCE 81

Figura 21. Ilustração da mensagem completa dada pela relação CONCESSION..... 84

Figura 22. Trecho da estrutura RST do texto CIENCIA_2000_6380 132

Figura 23. Trecho da estrutura RST do texto CIENCIA_2000_6381 133

Figura 24. Trecho da estrutura RST do texto CIENCIA_2000_17088 133

Figura 24. Trecho da estrutura RST do texto CIENCIA_2000_6381 134

Figura 25. Trecho da estrutura RST do texto CIENCIA_2000_6389 134

ÍNDICE DE TABELAS

Tabela 1. Classificação das Descrições Definidas..... 14

Tabela 2. Conjunto original das Relações RST 18

Tabela 3. Tipos de quebras de CCRs nos sumários produzidos pelo VeinSum..... 29

Tabela 4. Protótipos que compartilham os mesmos traços semânticos 45

Tabela 5. Quadro da correção da etiquetagem semântica para o *corpus* Summ-it..... 58

Tabela 6. Quadro de rebaixamento da saliência de EDUs 98

Tabela 7. Quadro de distância entre taxas de compressão 99

Tabela 8. Comparação da clareza referencial dos sumários do VeinSum com os gerados a partir desta proposta 100

ÍNDICE DE RESULTADOS INTERMEDIÁRIOS

ResInterm1. Resultados Intermediários do sumário CIENCIA_2001_6423..... 27

ResInterm2. Resultados Intermediários do sumário CIENCIA_2000_17101..... 30

ResInterm3. Resultados Intermediários do sumário CIENCIA_2003_24212.....31

ResInterm4. Resultados Intermediários do sumário CIENCIA_2000_17109.....32

ResInterm5. Resultados Intermediários do sumário CIENCIA_2000_6381.....33

ResInterm6. Resultados Intermediários do sumário CIENCIA_2000_17082.....34

ResInterm7. Resultados Intermediários do sumário CIENCIA_2000_17088.....35

ResInterm8. Resultados Intermediários do sumário CIENCIA_2000_17109.....36

ResInterm9. Resultados Intermediários do sumário CIENCIA_2000_17108.....37

ResInterm10. Resultados Intermediários do sumário CIENCIA_2001_17109.....38

ResInterm11. Resultados Intermediários do sumário CIENCIA_2001_19858.....73

ResInterm12. Resultados Intermediários do sumário CIENCIA_2000_17101.....77

ResInterm13. Resultados Intermediários do sumário CIENCIA_2000_17088.....82

ResInterm14. Resultados Intermediários do sumário CIENCIA_2000_17108.....85

ResInterm15. Resultados Intermediários do sumário CIENCIA_2000_17109.....87

ResInterm16. Resultados Intermediários do sumário CIENCIA_2000_6389.....90

ResInterm17. Resultados Intermediários do sumário CIENCIA_2000_17082.....91

ResInterm18. Resultados Intermediários do sumário CIENCIA_2000_6380.....93

ResInterm19. Resultados Intermediários do sumário CIENCIA_2000_6391.....95

ResInterm20. Resultados Intermediários do sumário CIENCIA_2000_6381.....96

ÍNDICE DE SUMÁRIOS[2]

Sum1. Sumário do texto CIENCIA_2001_6410 2

Sum2. Sumário do texto CIENCIA_2001_6423 26

Sum3. Sumário parcial do texto CIENCIA_2000_17082 29

Sum4. Sumário do texto CIENCIA_2000_17101 30

Sum5. Sumário do texto CIENCIA_2003_24212 31

Sum6. Sumário do texto CIENCIA_2000_17109 32

Sum7. Sumário do texto CIENCIA_2000_6381 33

Sum8. Sumário do texto CIENCIA_2000_17082 34

Sum9. Sumário do texto CIENCIA_2000_17088 35

Sum10. Sumário do texto CIENCIA_2000_17109 36

Sum11. Sumário do texto CIENCIA_2000_17108 37

Sum12. Sumário do texto CIENCIA_2001_17109 38

Sum13. Sumário do texto CIENCIA_2001_19858 73

Sum14. Sumário do texto CIENCIA_2000_17101 77

Sum15. Sumário do texto CIENCIA_2000_17088 82

Sum16. Sumário do texto CIENCIA_2000_17108 85

Sum17. Sumário do texto CIENCIA_2000_17109 87

[2] Todos os sumários constantes desta lista foram produzidos pelo VeinSum.

Sum18. Sumário do texto CIENCIA_2000_6389 90

Sum19. Sumário do texto CIENCIA_2000_17082 91

Sum20. Sumário do texto CIENCIA_2000_6380 92

Sum21. Sumário do texto CIENCIA_2000_6391 95

Sum22. Sumário do texto CIENCIA_2000_6381 96

ÍNDICE DE SUMÁRIOS NOVOS[3]

SumNovo1. Sumário novo do texto CIENCIA_2001_19858 74

SumNovo2. Sumário novo do texto CIENCIA_2000_17101 78

SumNovo3. Sumário novo do texto CIENCIA_2000_17088 83

SumNovo4. Sumário novo do texto CIENCIA_2000_17108 86

SumNovo5. Sumário novo do texto CIENCIA_2000_17109 88

SumNovo6. Sumário novo do texto CIENCIA_2000_6389 91

SumNovo7. Sumário novo do texto CIENCIA_2000_17082 92

SumNovo8. Sumário novo do texto CIENCIA_2000_6380 93

ÍNDICE DE TEXTOS-FONTE

TF1. Texto-fonte CIENCIA_2001_6410 2

TF2. Texto-fonte CIENCIA_2001_6423 22

[3] Os sumários constantes desta lista são os gerados a partir desta proposta de trabalho.

1 INTRODUÇÃO

A utilização de computadores com a finalidade de tratar a língua anos atrás era inconcebível para um linguista. Hoje em dia a Linguística Computacional ou Processamento das Línguas Naturais (PLN) é uma área de conhecimento em pleno desenvolvimento e desperta o interesse de inúmeros profissionais, dentre eles linguistas e informatas. Um dos principais objetivos da área é o desenvolvimento de sistemas capazes de interpretar, manipular e gerar mensagens codificadas em línguas naturais (Dias-da-Silva *et al.*, 2007).

Este projeto visa ao trabalho em uma subárea do PLN, a Sumarização Automática (SA) e estende as investigações do projeto ProCaCoSA[4], o qual teve por objetivo a construção de um sumarizador automático que privilegiasse o encadamento referencial dos sumários produzidos – o VeinSum (Carbonel, 2007).

Esse é um sumarizador de abordagem profunda e abrange três modelos que se complementam: a *Rhetorical Structure Theory*, ou RST (Mann & Thompson, 1988), a Teoria das Veias, ou VT (Cristea et al., 1998) e o Modelo de Saliência (Marcu, 1997). A RST é uma teoria que permite a organização do discurso em estruturas arbóreas interligadas por relações retóricas. A VT se baseia em uma árvore RST para determinar um conjunto de unidades de discurso que possa conter candidatos a antecedente, caso haja uma expressão anafórica na unidade textual selecionada para o sumário. No VeinSum, esse conjunto deve também ser incluído no sumário a cada unidade textual incluída, a fim de que se preservem os elos correferenciais. Para determinar quais unidades farão parte do sumário, o VeinSum obedece simultaneamante à VT e ao Modelo de Saliência, o qual torna obrigatória a inclusão de unidades textuais pela sua ordem de classificação de saliência, ou relevância, e não por sua interdependência referencial. Os três modelos são, portanto, complementares.

O problema é que não há qualquer processo de resolução anafórica no sistema, pois ele somente delimita contextos de possíveis antecedentes, sendo "cego" do ponto de vista linguístico. Isso pode gerar sumários que apresentam problemas de clareza referencial, como o constante em Sum1, o qual contém uma expressão anafórica sem antecedente explícito no

[4] PROcessamento de CAdeias de CO-referência em Sumarização Automática de Textos em Português do Brasil, Proc.CNPq Nro. 507030/2004-4, concluído em Outubro de 2008.

texto. Clareza referencial é definida como a propriedade de permitir ao leitor identificar a quem ou a que um determinado pronome ou sintagma nominal (SN) está se referindo[5].

Ao contrário do que muita gente pensa, a internet não está reduzindo os contatos entre as pessoas nem substituindo-os por relações impessoais conduzidas por computador. Segundo **o pesquisador**, os contatos via redes de computadores estão na verdade ampliando a socialização das pessoas.

Sum1. Sumário automático do texto CIENCIA_2001_6410[6]

Observa-se que não existe no sumário um antecedente para o SN **o pesquisador**, que no seu respectivo texto-fonte, ilustrado em TF1, aparece sublinhado. Deficiências dessa natureza comprometem a qualidade e a utilidade dos sumários, de forma geral, já que é impossível saber de quem se fala.

Ao contrário do que muita gente pensa, a internet não está reduzindo os contatos entre as pessoas nem substituindo-os por relações impessoais conduzidas por computador. A conclusão é de Barry Ellman, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá.
Segundo **o pesquisador**, os contatos via redes de computadores estão na verdade ampliando a socialização das pessoas.
Um dos exemplos que ele apresenta é o de um estudo feito em um subúrbio de Toronto, segundo o qual as pessoas "plugadas" em uma rede local conheciam três vezes mais vizinhos do que os não-conectados. Além disso, vizinhos conectados se encontraram pessoalmente 60% mais do que os excluídos da rede.
Os números gerais da internet apontam o mesmo fenômeno, diz Ellman. Segundo ele, pessoas ligadas por computadores tiveram mais contatos pessoais com seus amigos e parentes do que pessoas não-conectadas. O artigo do pesquisador está na edição de hoje da revista "Science".

TF1. Texto-fonte CIENCIA_2001_6410

Em casos como esse, o conjunto de possíveis antecedentes não contém o real antecedente da expressão anafórica e isso pode ocorrer devido a uma estruturação RST inadequada, pois a VT faz uso exclusivo de informações topológicas da árvore RST para determinar os conjuntos de possíveis antecedentes ou devido a erros no próprio algoritmo da VT ao calcular esse conjunto.

[5] Definição utilizada na DUC2005 - Document Understanding Conferences (http://duc.nist.gov/duc2005/).

[6] Todos os textos e sumários ilustrados neste trabalho são pertencentes ao *Corpus* Summ-it e são encontrados na íntegra no Apêndice A. O *corpus* está disponível para download em: (http://www.nilc.icmc.usp.br:8180/portal/index.jsp?option=downloads.jsp).

No entanto, considerando uma estrutura RST adequada, na maioria dos casos, os conjuntos de possíveis antecedentes contém esse real antecedente, além de unidades textuais que nada têm a ver com o contexto referencial da anáfora, ou seja, trechos de importância secundária no que tange à sua clareza referencial. Isso pode ser um problema que se agrava ainda mais sobre a restrição de compressão: o risco de violar esse parâmetro é alto devido à necessidade de inclusão do conjunto todo de possíveis antecedentes, além da própria unidade de informação saliente selecionada para compor o sumário. O VeinSum, então, a fim de respeitar ao máximo a taxa de compressão (TC), despreza unidades textuais mais salientes em prol de unidades cujos conjuntos de possíveis antecedentes sejam menores. Define-se TC como a porcentagem do texto-fonte que se preserva no sumário, conforme Mani (2001). Neste trabalho, a TC adotada regularmente é de 30%.

A fim de evitar o descarte de unidades salientes, o objetivo principal deste trabalho é a proposta de um modelo heurístico que utiliza conhecimento semântico proveniente do *parser* PALAVRAS (Bick, 2000), para complementar os modelos de decisão utilizados pelo VeinSum no processo de seleção de unidades textuais relevantes (Tomazela & Rino, 2009). Mais especificamente, utilizar as etiquetas semânticas do PALAVRAS para a identificação de unidades textuais correferentes e ao invés de ter um conjunto com vários possíveis antecedentes, tentar apontar o correto. Isso deve permitir o descarte das demais unidades textuais do conjunto e esse passará a conter somente dois segmentos textuais: o da anáfora e o do seu antecedente. Assim, são maiores as chances de o sumário conter segmentos salientes, indicados pelo Modelo de Marcu, já que a probabilidade de o VeinSum ter que descartar unidades salientes que têm o conjunto de possíveis antecedentes extenso demais, em prol de outras menos salientes mas que tenham esse conjunto mais reduzido, é menor.

Para que essa identificação automática de termos correferentes fosse possível, métodos de aprendizado de máquina foram utilizados na geração de heurísticas de seleção e filtragem de segmentos correferentes. Essas heurísticas consideram as etiquetas semânticas atribuídas aos núcleos dos SNs que compõem os 50 textos jornalísticos do *Corpus* Summ-it (Collovini et al., 2007) e determinam que, para um segmento anafórico incluído no sumário, o segmento que contém o seu antecedente deve também ser incluído e tão somente ele, desprezando as demais unidades.

Para o desenvolvimento deste trabalho, partiu-se das seguintes premissas: (i) as anotações RST e de correferência para o *Corpus* Summ-it, feitas manualmente por especialistas, são adequadas; (ii) os conjuntos de possíveis antecedentes de um segmento

textual anafórico, determinados pela VT, contêm segmentos realmente correferentes a ele; (iii) o conjunto de possíveis antecedentes de um segmento anafórico contém o seu antecedente mais completo.

As hipóteses estabelecidas no início do trabalho foram que: (i) os elementos textuais correferentes possuem etiquetas semânticas iguais ou similares; (ii) a semântica do PALAVRAS é suficientemente delineada para identificar unidades textuais possivelmente correferentes (iii) a aplicação das heurísticas possibilita a redução do conjunto de candidatos a antecedente, os quais são indicados automaticamente; (iv) a aplicação das heurísticas não compromete o encadeamento referencial do texto-fonte; (v) a aplicação das heurísticas permite a manutenção da clareza referencial; (vi) a aplicação das heurísticas em sumários que corrompem a TC pode reduzir o tamanho dos mesmos e, consequentemente, fazer com que a TC se aproxime mais da considerada ideal; (vii) com a redução do conjunto de possíveis antecedentes, o VeinSum deixará de descartar contextos referenciais salientes e passará a respeitar melhor a saliência das unidades de discurso, para a inclusão em um sumário.

De forma geral, a proposta de uso de informações semânticas como complemento dos recursos já utilizados pelo sumarizador prova ser eficiente, porém somente algumas das hipóteses acima enumeradas foram comprovadas e uma das premissas também se mostrou inadequada.

Alguns problemas de desempenho do VeinSum sugerem que as anotações feitas manualmente por especialistas contém alguns erros e devem ser revistas e a premissa (i) não pôde ser comprovada. Ainda, a hipótese (iv) não se mostrou adequada, já que a utilização das heurísticas corrompeu um caso de encadeamento referencial do texto-fonte. De início a anotação semântica do PALAVRAS mostrou-se insuficiente para indicar entidades correferentes devido ao grande número de inconsistências apresentadas, porém, ao invés de mudar os rumos deste trabalho, optou-se por uma pós-edição manual criteriosa, corrigindo as inconsistências da anotação semântica para o *corpus* Summ-it. A partir dessa pós-edição manual, a anotação semântica mostrou-se suficiente para atingir o objetivo desta proposta e a hipótese (ii) pôde ser comprovada, portanto.

Como contribuições principais deste trabalho tem-se a proposta do novo modelo de SA, com heurísticas que refinam o processo de reconhecimento de elementos textuais correferentes, além das avaliações críticas ao desempenho do VeinSum e do *parser* PALAVRAS, seguida da posterior pós-edição manual de sua anotação semântica para o *corpus*

Summ-it inteiro, o que possibilita que outras pesquisas na área de PLN façam uso dessa anotação de forma mais eficiente.

Nesta monografia, é descrito, no capítulo 2, o processo de sumarização automática com foco na abordagem profunda. No capítulo 3, são apresentados os conceitos linguísticos pertinentes ao estudo desenvolvido neste trabalho. O capítulo 4 contém uma descrição dos modelos utilizados no sumarizador e, no capítulo 5, são descritos o detalhamento do funcionamento do VeinSum, além dos principais problemas de desempenho apresentados por ele. O capítulo 6 contém a descrição detalhada desta pesquisa e também os sumários gerados a partir desta proposta, seguidos de análises linguísticas que indicam as vantagens e fragilidades de se utilizar tal abordagem. Finalmente, no capítulo 7, são apresentadas as considerações finais desta pesquisa.

2 A SUMARIZAÇÃO AUTOMÁTICA

O Veinsum é um sumarizador que utiliza um alto nível de conhecimento linguístico para a construção dos sumários e segue a abordagem de SA considerada profunda. Devido a isso, esse capítulo é dedicado inteiramente ao esclarecimento de conceitos relacionados à SA e somente a abordagem profunda é considerada.

Segundo Mani (2001), um sumarizador é um sistema cujo objetivo é 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.

As principais premissas da sumarização podem ser numeradas como segue (Rino & Pardo, 2003, p. 2):

- 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 ideia central – o tópico principal do texto – sobre a qual se constrói a trama textual;
    b) um conjunto de unidades de informação que, reconhecidamente, têm relação com a ideia central em desenvolvimento;
    c) um objetivo comunicativo central que, implícita ou explicitamente, direciona tanto a seleção das unidades de informação quanto o modo como a informação será estruturada, para estabelecer a ideia pretendida;
    d) um enredo, tecido em função das escolhas antes citadas, visando transmitir a ideia 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 ideia central, sem transgredir o significado original pretendido.
- A não transgressão do original constitui a principal restrição da sumarização.

As etapas acima descritas são realizadas mentalmente através dos processos cognitivos aos quais o humano se submete durante o processo de sumarização. Esse, ao condensar o

conteúdo de um texto, faz uso de certo grau de conhecimento linguístico a fim de reformular estruturas e fazer suas escolhas lexicais. Ferramentas baseadas nessa abordagem buscam simular, analogamente, o processo humano de sumarização através de tarefas algorítmicas. Sumarizadores com essa abordagem possuem uma arquitetura refinada de processos automáticos, com modelos linguísticos e/ou discursivos, os quais envolvem um alto nível de conhecimento profundo.

A Figura 1, adaptada de Mani & Maybury (1999, p. ix), descreve a arquitetura de um sistema de abordagem profunda. O sistema faz a análise das informações, reconhecendo dados relevantes e eliminando informações desnecessárias, seguida da síntese das unidades textuais essenciais, o que resulta em um sumário.

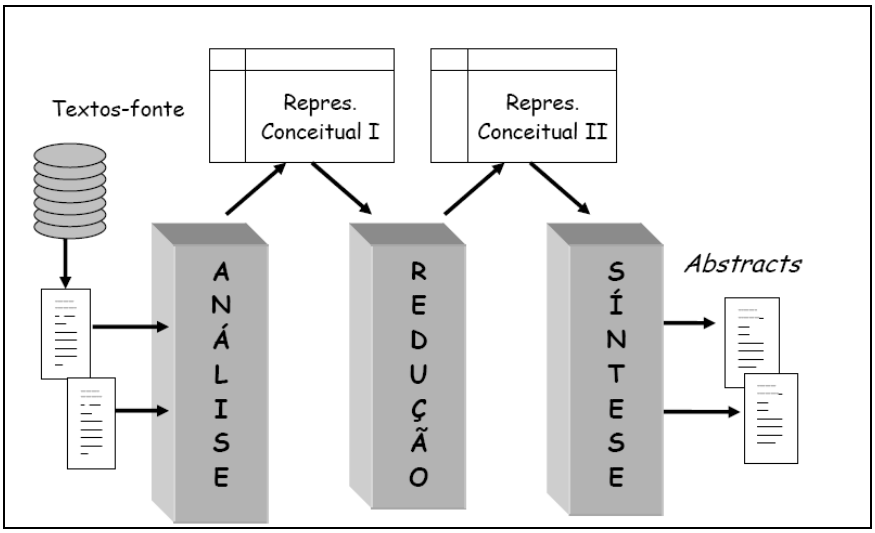


Figura 1. Arquitetura geral de um sistema de SA de abordagem profunda

A divisão da SA em três partes distintas foi proposta por Sparck Jones (1993) e é organizada da seguinte forma:

- Texto-fonte é transformado em uma representação computável (representação conceitual I);
- Redução dessa primeira representação; ainda em formato não-textual (representação conceitual II);
- Realização linguística, ou reescrita textual, da representação conceitual condensada.

O processo de análise corresponde à interpretação do texto-fonte, gerando, assim, uma representação do conhecimento linguístico expresso em termos computáveis. A redução é a etapa em que esse conhecimento é manipulado, gerando o sumário ainda sob a forma de uma representação computável. E na tarefa de síntese, por fim, o sumário ganha a forma textual.

É importante salientar que, na fase de realização linguística, as escolhas lexicais e morfossintáticas podem ser diferentes das constantes no texto-fonte, o que dependerá do refinamento de conhecimento linguístico disponível no sistema.

Apesar da maior dificuldade em se construir sistemas de abordagem profunda, devido à complexidade de simulação do processo cognitivo e do fato de os sistemas já existentes ainda apresentarem alguns problemas de textualidade, os resultados obtidos com ferramentas que utilizam essa abordagem apresentam-se cada dia mais satisfatórios, como: Marcu (1997), (1999), (2000), Azzam et al., (1999), Pardo (2002), Cristea (2005), Seno (2005); Carbonel (2007); Jorge & Pardo (2010), Jorge (2010), Uzêda *et al.* (2009) Uzêda *et al*. (2010).

Dentre os problemas de textualidade que um sumário pode apresentar estão os relacionados à clareza referencial, também foco deste trabalho. No próximo capítulo, encontram-se detalhadamente descritos o conceito de textualidade, assim como os fatores que a englobam intratextualmente, como a coerência e a coesão.

3 CONCEITOS LINGUÍSTICOS RELEVANTES PARA ESTA PESQUISA

Como um dos focos deste trabalho é a manutenção da textualidade no que tange à clareza referencial dos sumários gerados pelo VeinSum, é necessário considerar conceitos básicos de Linguística Textual.

A Linguística Textual ou Linguística do Texto (LT) é a área da Linguística que trata o texto como objeto de investigação. Surgiu na Europa, na década de 60 e somente a partir dos anos 80 é que as Teorias do Texto ganharam maior prestígio.

Na fase inicial da LT, falava-se em texto (constituintes linguísticos coerentemente organizados) e não texto (constituintes linguísticos organizados sem coerência). Com os avanços nos estudos da área, elementos que se relacionam diretamente com a concepção (pelo falante) e recepção do texto (por um interlocutor) passam a ser estudados e, não somente, os elementos de sua interioridade.

O texto passa, então, a ser um todo muito maior que a simples soma de frases e palavras, pois existem fenômenos linguísticos que podem somente ser explicados dentro do todo. Deve preservar a coesão superficial, ou seja, a organização linear ao nível dos constituintes linguísticos, além de a organização reticulada ou tentacular, não linear, portanto, dos níveis de sentido e intenções que garantem a coerência no aspecto semântico e funções pragmáticas, como as pressuposições e as implicações consideradas na produção do sentido. (Marcuschi, 1983).

Como se pode observar, aspectos como coerência e coesão são considerados na produção textual e contribuem com a boa organização do texto, porém não são os únicos responsáveis quando se trata da textualidade, conceito descrito na próxima seção.

3.1 Textualidade

O conceito de textualidade é formado por vários critérios intra e extralinguísticos, como definem Beaugrande & Dressler (1981). A coesão e a coerência são os fatores intralinguísticos (centrados no texto) e a informatividade, a situacionalidade, a intertextualidade, a intencionalidade e a aceitabilidade são considerados fatores extralinguísticos (centrados no usuário).

Nesta pesquisa, entretanto, a textualidade é somente tratada no nível intratextual, ou seja, somente aspectos como coesão e coerência são considerados. Esses conceitos são descritos, respectivamente, nas seções 3.1.1 e 3.1.2.

3.1.1 Coerência Textual

No início dos estudos da LT, os conceitos de coesão e coerência eram, muitas vezes, equiparados um ao outro. Com o avanço dos estudos nessa área, houve uma mudança gradual no conceito de coerência, a qual deixa de ser considerada uma simples propriedade do texto e passa a ser um fenômeno mais amplo. De início, a coerência foi diferenciada da coesão e abrangia aspectos sintáticos e semânticos, ou seja, levava em consideração apenas a superfície textual. Descobriu-se, no entanto, que na superfície textual encontrava-se apenas parte do sentido de um texto, mas nunca a totalidade de suas informações semânticas (Koch, 2004a).

Assim, os estudiosos do texto foram além dessa perspectiva e ele passa a ser visto como instrumento de realização de intenções comunicativas do falante, ou seja, o ouvinte não se limita a "entender" o texto, mas a reconstruir os propósitos comunicativos que tem o falante ao estruturá-lo, o qual faz suas escolhas sintáticas e semânticas a partir de sua intenção comunicativa. Considera-se, então, uma perspectiva pragmático-enunciativa, ou seja, a coerência se constrói da interação entre o texto e seus usuários.

Ela deve ser entendida como um princípio de interpretabilidade, ligada à inteligibilidade do texto numa situação de comunicação (Koch & Travaglia, 2004). Tem a ver com a boa formação do texto num sentido totalmente diverso da noção de gramaticalidade, mas relaciona-se à organização subjacente do mesmo, embora os elementos da superfície linguística possam servir de pista para o seu estabelecimento.

Para Halliday & Hasan (1976), a coerência se constrói em um nível que ultrapassa os traços linguísticos do texto; articula elementos não somente de ordem linguística, mas também cognitiva e interacional; não basta que haja conectividade linguística entre os segmentos do texto, é preciso que haja relações de sentido entre os mesmos. Para esses autores, a coesão e a coerência estão intimamente ligadas, uma vez que a segunda depende da adequação da primeira.

No entanto, para Marcuschi (1983) e Koch & Travaglia (2004), os conceitos de coesão e coerência são independentes. Para eles, a continuidade se dá ao nível do sentido e não ao nível das relações entre os constituintes linguísticos, ou seja, podem existir textos que não

apresentem elos coesivos, porém que sejam coerentes em seu todo, ou textos coesos, mas totalmente desprovidos de sentido, respectivamente ilustrados pelos casos abaixo[7]:

(1) "Alegria no olhar. Desejo de liberdade. Os portões se abrem. A família a espera. Abraços. Choro. Esperança. Vida nova."

(2) "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 desligou o rádio."

No caso (1), o encadeamento dos enunciados é totalmente desprovido de elos coesivos explícitos. Há somente uma justaposição de informações distintas, sem uma estrutura elaborada, porém é possível resgatar uma construção progressiva de sentido. O encadeamento coesivo se dá pela progressão semântica do texto. No caso (2), tem-se uma sequência de enunciados encadeados por recursos coesivos, mais especificamente pelo uso de marcadores de discurso, porém isoladamente, eles não são suficientes para permitir o resgate de algum sentido desse enunciado.

É preciso salientar, entretanto, que no escopo deste trabalho, o tratamento da coerência limita-se às contribuições que a boa estruturação do texto fornece à sua construção de sentido. O aspecto pragmático da coerência textual não será abordado neste trabalho.

Na próxima seção, tem-se uma descrição mais detalhada sobre o segundo tópico de textualidade que envolve aspectos intratextuais – a coesão.

3.1.2 Coesão Textual

Para que um texto seja coerente, ele precisa articular de modo eficiente seus constituintes linguísticos a fim de ter sua mensagem depreendida. O encadeamento desses constituintes é o que forma a tessitura, o tecido textual (Koch, 2004b), ou ainda chamada, coesão. A coesão é explicitamente marcada através de elementos linguísticos, o que lhe dá um caráter linear, ao contrário da coerência.

Para Halliday & Hasan (1976), em sua obra clássica sobre coesão textual, essa ocorre quando a interpretação de algum elemento no discurso é dependente da de outro. Um pressupõe o outro, no sentido que, esse não pode ser efetivamente decodificado a não ser por

[7] Textos (1) e (2) foram criados para ilustrar os casos.

recurso ao outro. A cada ocorrência de um recurso coesivo no texto, denominam laço ou elo coesivo, que, segundo eles, são os principais fatores de coesão.

Os elementos coesivos dão maior legibilidade ao texto, reduzindo a possibilidade de leituras equivocadas e a ausência desses elementos estruturais, além de empobrecer o texto, pode também implicar na impossibilidade de entendimento e na sua má estruturação, o que, no caso da SA, compromete a coerência dos sumários.

Um dos principais fatores de coesão para Halliday e Hasan (1976) é a referência, que pode ser endofórica, quando o referente se acha expresso no próprio texto, ou exofórica, quando há remissão a algum elemento da situação comunicativa. Unindo-se os conceitos de coesão e referência, temos a coesão referencial, que é definida, segundo Koch (2004a, p.31), como "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"; o primeiro denomina-se forma referencial e o segundo referente textual. Fala-se em remissão de um componente do texto a outro nele presente quando se pode apontar, no texto, uma entidade concreta que é definida como antecedente da expressão referencial. No segundo caso descrito, a inferência é feita a partir do contexto, ou seja, o elemento de referência não se encontra no texto. Neste trabalho, no entanto, contempla-se exclusivamente o caso da referência endofórica.

A coesão referencial pode ser obtida por meio de dois mecanismos básicos (Koch & Travaglia, 2004, p. 48):

(i) **substituição:** quando um componente da superfície textual é retomado (anáfora) ou precedido (catáfora) por uma pró-forma (pronome, verbo, advérbio, quantificadores que substituem outros elementos do texto). Há também a substituição por zero, que é a elipse.

(ii) **reiteração:** que se faz através de sinônimos, de hiperônimos, de nomes genéricos, de expressões nominais definidas, de repetição do mesmo item lexical, de nominalizações.

Esse tipo de coesão é marcado no texto através de Cadeias de Correferência (CCRs), as quais são definidas na próxima seção.

3.2 Cadeias de Correferência

Dentre os vários mecanismos coesivos presentes em um texto estão as CCRs, que expressam a coesão referencial, como no exemplo extraído de Koch (2004a, p.32):

- Ontem fui conhecer a nova casa de Alice. Ela é moderna e bem decorada.

No exemplo, pode-se afirmar que o pronome Ela se refere à nova casa de Alice, pois ao estabelecer conexão com o seu antecedente, encontra sentido.

No entanto, no exemplo que segue, o caso é diferente, apesar de a primeira oração ser idêntica à do exemplo anterior:

- Ontem fui conhecer a nova casa de Alice. Ela a comprou com a herança recebida dos pais.

Nesse caso, o SN a nova casa de Alice e o pronome Ela não possuem o mesmo referente, pois fazem remissão a entidades distintas no mundo real: a nova casa de Alice e Alice, respectivamente, não sendo correferentes, portanto.

O fenômeno da correferenciação, então, se manifesta quando há uma "identidade de referência" entre elementos textuais, ou seja, quando dois ou mais itens lexicais expressam o mesmo referente. Quase todos os estudos sobre coesão referencial partem do pressuposto de que existe identidade de referência entre a forma remissiva e seu referente. Tal identidade, porém, é discutível, pois o referente pode ser acrescido de outros traços que se lhe vão agregando ao longo do texto.

No escopo deste trabalho, no entanto, apenas uma fração do total de fenômenos referenciais é tratada; exclusivamente as descrições definidas (DDs) em contextos correferenciais. As DDs são SNs antecedidos de artigo definido, como nos trechos grifados abaixo:

O presidente Lula, na última semana, assinou o acordo histórico que impulsionará a produção de biodiesel no Brasil.

Segundo Vieira (1998) e Coelho *et al* (2006), as DDs são divididas em quatro classes, dependendo de como estão relacionadas com os seus antecedentes. Se forem correferentes, são chamadas anáforas diretas e indiretas. Entende-se por anáfora direta aquela que retoma seu antecedente através do mesmo nome núcleo e anáfora indireta aquela que o retoma através de nome núcleo diferente, o qual pode ser expresso por sinonímia, hiperonímia ou hiponímia. Se

não forem correferentes, são consideradas anáforas associativas ou formas não anafóricas. Uma anáfora associativa introduz um novo referente que possui parte de seu significado ancorado em uma expressão mencionada anteriormente a ela e as formas não anafóricas introduzem um novo referente que não tem seu sentido ancorado em nenhuma outra expressão do texto.

A síntese desta classificação consta da Tabela 1, adaptada de Carbonel (2007, p.36).

Tabela 1. Classificação das Descrições Definidas

DDs	**Correferentes**	**Anáforas diretas**	**O presidente** viajou para o exterior. **O presidente** levou consigo uma grande comitiva.
		Anáforas indiretas	O professor mostrou **a prova** aos alunos. **O teste** já estava corrigido.
	Não Correferentes	**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)

Como o foco deste trabalho encontra-se em CCRs, somente são consideradas as anáforas que indicam elementos textuais correferentes, ou seja, somente anáforas diretas e indiretas.

Vale destacar que elementos linguísticos correferentes podem ou não preservar suas características morfológicas, como nos exemplos abaixo:

Casos em que as características morfológicas são preservadas:

- Gênero é mantido - o imóvel e o apartamento;
- Número é mantido – os estudantes e os alunos.

Casos em que as características morfológicas são variadas:

- Mudança de gênero - a prova e o teste;
- Mudança de número – os estudantes e o corpo discente.

A capacidade de reconhecimento dessa relação por um humano é trivial, ao passo que para a modelagem computacional constitui uma limitação, pois somente baseado em características morfossintáticas não se pode fazer o reconhecimento automático de itens lexicais como os acima citados. Isso acaba sendo um dos fatores mais problemáticos na construção de sumários, pois pode comprometer sua coerência. Devido a isso, propõe-se a utilização de etiquetas semânticas no processo de reconhecimento automático dessas entidades.

Encontram-se, no capítulo seguinte, as formas de tratamento computacional do texto que exploram técnicas utilizadas para o tratamento computacional de CCRs.

4 PRESSUPOSTOS TEÓRICOS DO PROJETO

Neste capítulo, encontra-se um apanhado geral dos modelos teóricos que fornecem subsídios para a manipulação da informação textual e que são úteis também para modelagens que privilegiam o tratamento de CCRs.

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

A RST, do inglês *Rhetorical Structure Theory*, é uma teoria que permite a organização do discurso em estruturas RST (também chamadas árvores RST devido à sua estrutura arbórea), as quais são construídas gradualmente pela junção de suas unidades elementares, ou EDUs (do inglês, *Elementary Discourse Units*), através de relações retóricas, ou relações RST.

Tais relações indicam os tipos de ligação existentes entre duas proposições, normalmente expressas por segmentos adjacentes, visando à organização coerente do texto. A cada EDU é atribuído um papel de núcleo (N) ou de satélite (S), dependendo do grau de importância que tal unidade expressa no texto, isto é, N expressa informação mais relevante comparada ao conteúdo de S, que apresenta informação complementar. Opondo-se às relações mononucleares, que possuem N e S, existem casos em que ambas as unidades trazem informação igualmente importantes, chamadas de relações multinucleares, que possuem mais de um N e nenhum S. Na Figura 2, encontra-se um exemplo de relação mononuclear e, na Figura 3, um exemplo de relação multinuclear.

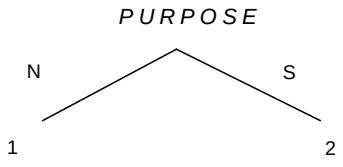
Durante os últimos oito anos, uma rede de satélites do Departamento de Defesa dos EUA tem monitorado a atmosfera terrestre [1] com o objetivo de detectar explosões [2].	

Figura 2. Excerto e estrutura RST que evidenciam uma relação mononuclear

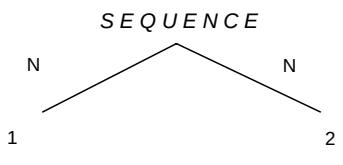
A larva passa de 7 a 14 dias ali dentro [1] e, então, libera uma droga ainda desconhecida na corrente sanguínea da vítima [2].	

Figura 3. Excerto e estrutura RST que evidenciam uma relação multinuclear

A definição de cada relação especifica as informações necessárias para que se faça a escolha de uma em detrimento da outra. Tais informações são organizadas como segue:

- Restrições sobre o núcleo (N);
- Restrições sobre o satélite (S);
- Restrições sobre a determinação de núcleo ou de satélite (N+S)
- Efeito: especificação do efeito que o escritor pretende causar no leitor com a escolha de tal relação retórica.

A seguir, encontram-se as definições das relações PURPOSE e SEQUENCE, utilizadas nos exemplos acima, traduzidas por Pardo (2005) e apresentadas nas Figura 4 e Figura 5, respectivamente.

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 conhece que a atividade em N pode ser iniciada por meio de S

Figura 4. 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 5. Definição da relação SEQUENCE

O conjunto original totaliza 24 relações retóricas e é ilustrado na Tabela 2**.** Neste trabalho, os nomes originais das relações retóricas, em inglês, são mantidos, haja vista a prática comum em outros trabalhos na área de Linguística Computacional.

Tabela 2. Conjunto original das Relações RST

Relação	Mono-nuclear	Multi-nuclear	Relação	Mono-nuclear	Multi-nuclear
ANTITHESIS	**x**		*JUSTIFY*	**x**	
BACKGROUND	**x**		*MOTIVATION*	**x**	
CIRCUMSTANCE	**x**		*NON-VOLITIONAL 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**	

Trabalhos posteriores ao de Mann & Thompson propõem conjuntos bastante extensos, contando com mais de cem relações, o que torna a tarefa de construção da estrutura arbórea do texto alvo muito mais complexa. Algumas das relações retóricas previstas no conjunto de Marcu (1997) são bastante intuitivas para um falante competente, ao passo que outras são mais obscuras, forçando o analista a recorrer à obra de referência a fim de entender suas definições.

Neste trabalho, além das 24 relações retóricas originais, são utilizadas 8 adotadas por Pardo (2005) ao propor a análise automática para o português com o DiZer, ferramenta considerada no Projeto ProCaCoSA, as quais são: ATTRIBUTION, COMPARISON, CONCLUSION, EXPLANATION, MEANS, PARENTHETICAL, SAME-UNIT E LIST, totalizando 32 relações retóricas.

Após a utilização dessas relações para a construção de subárvores simples, essas são ligadas a outras subárvores simples por um processo composicional até que o texto esteja inteiramente interconectado. Observa-se, na Figura 6, um excerto do texto CIENCIA_2000_6391, segmentado de acordo com a teoria e, na Figura 7, a estrutura RST correspondente, feita por um especialista.

[1] Sete ativistas do Greenpeace foram presos ontem nos EUA [2] ao tentar impedir o descarregamento de madeira brasileira na cidade de Savannah, na costa do estado da Georgia. [3] Os ativistas [4] (norte-americanos e europeus) [5] foram rechaçados pelos tripulantes de um navio de bandeira dinamarquesa e presos pela polícia.

Figura 6. Excerto do texto CIENCIA_2000_6391

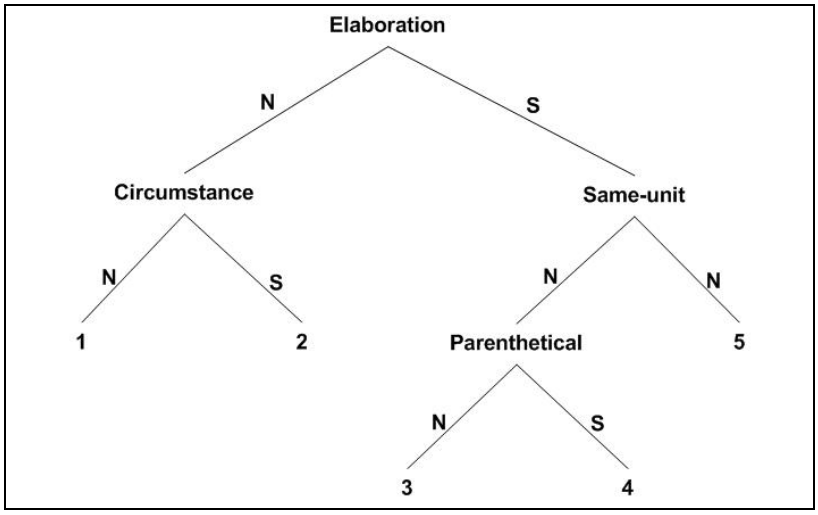


Figura 7. Árvore RST do excerto da Figura 6

Como a tarefa de análise baseia-se na intenção subjacente ao texto e a tarefa de interpretação é extremamente subjetiva, os próprios autores da teoria reconhecem que, a partir de um único texto pode-se obter mais de uma estrutura arbórea. Tal subjetividade pode gerar discordância entre anotadores especialistas ou ainda discrepâncias de anotação advindas de um mesmo anotador. As variações de anotação podem ocorrer em diversos níveis de análise, como por exemplo: (i) na delimitação das unidades mínimas de significado; (ii) na nomeação de N e S; (iii) na escolha das relações retóricas utilizadas para relacionar as proposições, e ainda (iv) na sua estrutura retórica em si, devido a possíveis interpretações diferentes da mensagem principal expressa pelo texto-fonte.

Conforme já propuseram diversos sistemas de SA (p.ex: (Marcu, 1997), (Marcu, 1999), (Marcu, 2000), (O'Donnell, 1997), (Ono et al., 1994)) são as relações mononucleares

que permitem delinear EDUs supérfluas e, assim, candidatas à exclusão de um sumário[8] (Sparck Jones, 1993). Marcu parte dessa ideia para considerar que um N expressa informação mais saliente, quando comparado ao seu S, cuja exclusão não prejudicaria a qualidade do sumário. O problema de se construir sumarizadores baseados somente nessa ideia de nuclearidade da RST é que, muitas vezes, as informações constantes em Ss não são supérfluas. Poderia ocorrer um antecedente de uma anáfora em um S, por exemplo, que obrigasse sua escolha para o sumário. Assim, além de não se poder excluir indiscriminadamente os Ss, fica impossível tratar o fenômeno do encadeamento referencial, contemplado aqui pela utilização da Teoria das Veias, descrita a seguir.

4.2 *Veins Theory* (VT) - Teoria das Veias

A VT é uma teoria que se define sobre estruturas RST e, portanto, depende da construção prévia das mesmas. Ela propõe um modelo que visa garantir que, a partir da coerência local, um discurso seja coerente em seu todo e é incorporada ao modelo de SA para evitar que haja problemas de clareza referencial no sumário.

Suas regras pretendem determinar, a partir da topologia de uma árvore RST, a veia e o chamado domínio de acessibilidade referencial (*acc*) para cada unidade do discurso. A veia de uma EDU consiste em um conjunto de EDUs que podem conter anáforas e catáforas relacionadas a uma entidade; em outras palavras, abrange, supostamente, os elementos encadeáveis referencialmente à EDU em foco. Já o *acc* corresponde ao conjunto de EDUs inseridas na veia que estão antepostas à EDU em foco, ou seja, engloba somente as anáforas e é um subconjunto da veia. Essa noção de veia decorre de observações de como as referências se comportam em uma árvore RST.

Segundo Cristea (2003; 2005), considerando a organização hierárquica e o princípio de composicionalidade, o qual permite relações de longa distância entre nós parentes, isto é, que EDUs distantes umas das outras na superfície textual sejam interligadas por relações RST, as intuições da teoria determinam que:

[8] Afinal, é impossível distinguir diferentes graus de importância para EDUs multinucleares, por isso, a inclusão de todas elas é considerada.

1) satélites ou núcleos à direita podem se referir a núcleos de mesmo nível (irmãos, portanto) à esquerda;

2) um núcleo à direita pode se referir a um satélite à esquerda;

3) um satélite à direita de um núcleo N não é acessível de um outro irmão à direita de N, seja ele um núcleo ou um satélite;

4) um núcleo bloqueia a referência de um satélite à direita para um à esquerda.

A ideia aqui é que a distinção entre núcleo e satélite restringe a gama de referentes pelos quais as anáforas podem ser resolvidas e, assim, para uma EDU que contém uma anáfora, Cristea afirma que seu antecedente será certamente encontrado em seu *acc*, salvo em casos de difícil solução na superfície textual.

Na Figura 8, pode-se visualizar a representação RST do excerto da Figura 6 anotado com o *acc* de cada EDU. O texto está novamente reproduzido abaixo:

> [1] **Sete ativistas do Greenpeace** foram presos ontem nos EUA [2] ao tentar impedir o descarregamento de madeira brasileira na cidade de Savannah, na costa do estado da Georgia. [3] Os ativistas [4] (norte-americanos e europeus) [5] foram rechaçados pelos tripulantes de um navio de bandeira dinamarquesa e presos pela polícia.

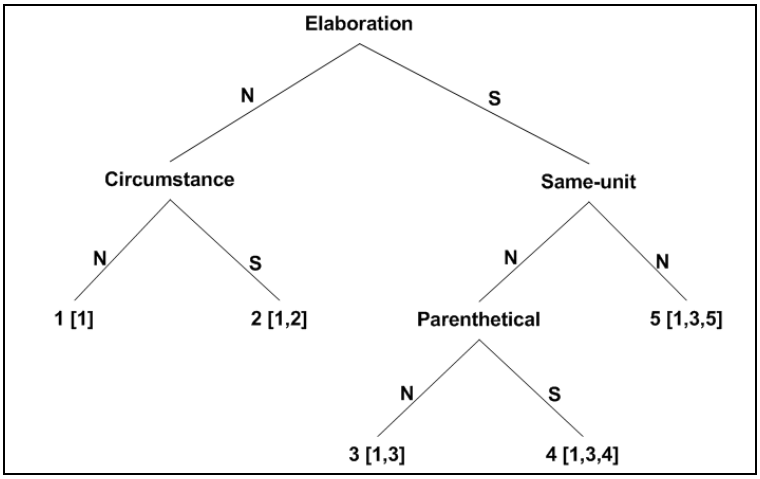


Figura 8. Árvore RST anterior anotada com o *acc* de cada EDU

Nesse texto, a anáfora os ativistas, grifada na EDU 3, remete diretamente ao seu antecedente sete ativistas do Greenpeace, em negrito na EDU 1. Pelo *acc*(3)={1,3}, verifica-se que a teoria aponta, acertadamente, para a EDU que contém o antecedente. Esse cálculo, no entanto, nem sempre é totalmente eficiente, pois além de, muitas vezes, incluir EDUs que não

fazem parte do contexto referencial em foco, há casos em que a EDU que contém o antecedente de uma expressão anafórica não está inserida em seu *acc*.

Devido a isso, entende-se que a teoria não é suficiente para a correta identificação de termos correferentes e outros recursos devem ser considerados para evitar que o sistema produza sumários com problemas de encadeamento referencial ou com EDUs de importância secundária.

A seguir, encontra-se descrito o modelo prático que determina como as EDUs são escolhidas para compor o sumário, no VeinSum.

4.3 Modelo de Saliência

Assim como a VT, o Modelo de Saliência também se baseia em uma árvore RST, porém para calcular e atribuir pesos a todas as suas EDUs, os quais refletem sua saliência na árvore.

A fim de ilustrar como o cálculo é realizado, encontra-se abaixo o texto CIENCIA_2001_6423 (com sua marcação de EDUs), ilustrado em TF2, seguido de sua árvore RST, constante da Figura 9.

[1] O mecanismo que faz as pessoas sentirem falta de ar em regiões montanhosas ou depois de uma corrida está fortemente ligado a gases da família do óxido nítrico, [2] de acordo com estudo publicado na última edição da revista "Nature" [3] (www.nature.com). [4] Cientistas da Universidade da Virgínia, EUA, descobriram que esses gases, [5] chamados S-nitrosotióis [6] (Snos) [7] atuam em todos os níveis da regulação respiratória, [8] fazendo com que os vasos sanguíneos e vias respiratórias dilatem, [9] sinalizando a necessidade de oxigênio por parte dos tecidos [10] e comunicando-se com as regiões cerebrais que controlam o desejo de respirar. [11] Segundo Benjamin Gaston, um dos cientistas envolvidos na pesquisa, [12] a descoberta pode abrir alternativas para o tratamento de disfunções respiratórias. [13] Stuart Lipton, do Instituto Burnham, Califórnia, não duvida: [14] "Esse estudo tem lugar garantido nos manuais de fisiologia", [15] diz.

TF2. Texto-fonte CIENCIA_2001_6423

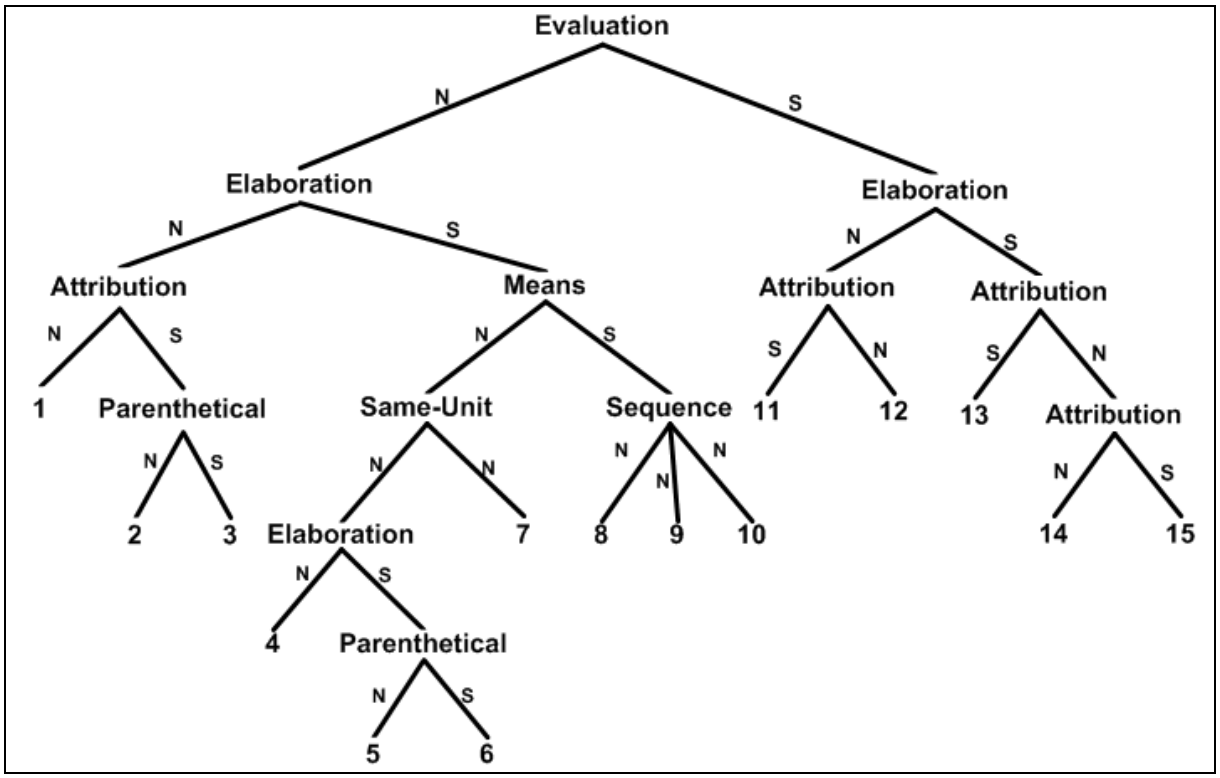


Figura 9. Estrutura RST do texto CIENCIA_2001_6423

Segundo Marcu, as EDUs nucleares que estão mais próximas da raiz da árvore têm peso maior e são, portanto, mais salientes que as EDUs que se encontram em níveis mais profundos. Assim, o cômputo da saliência das EDUs se baseia tanto na nuclearidade quanto em sua profundidade na estrutura RST. Essa determinação ocorre de maneira *bottom-up*, como segue:

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

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

$$s(u,D,d) = \begin{cases} 0 & \text{se } D \text{ é nulo,} \\ d & \text{se } u \in \text{prom(D),} \\ d\text{-}1 & \text{se } u \in \textit{parentheticals} \text{ (D),} \\ \textit{max}(\textit{score}\ (u, \textit{leftChild}(\text{D}), \text{d-1}), & \\ \quad \textit{score}\ (u, \textit{rightChild}(\text{D}), \text{d-1})) & \text{caso contrário} \end{cases}$$

em que:

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

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

Para qualquer EDU *u* de uma árvore RST, seu conjunto promocional consiste de um conjunto formado unicamente por sua própria unidade promocional, ou seja, conjunto promocional(u)={u}. Neste caso, *u* é o único nó e não compreende qualquer relação RST da árvore.

Para qualquer unidade promocional, o conjunto promocional de *D*, sendo *D* uma árvore RST, será: conjunto promocional(*D*)={u_1,u_2,...u_n}, onde u_i, sendo que *i* varia de 1 a *n*, é uma unidade promocional de *D*.

Ainda, o conjunto promocional de *D* é constituído pelo N mais nuclear e é indicado na sua raiz. Esse conjunto será unitário, caso as relações RST envolvidas com posições nucleares não levarem a relações multinucleares. Caso essas sejam consideradas, o conjunto promocional de *D* será um conjunto com mais de um elemento.

Aplicando-se repetidamente esse algoritmo a cada nó da árvore RST, pode-se obter a ordem de saliência de todas as EDUs do texto-fonte. Para a estrutura RST apresentada acima, tem-se: 1>12>4,7,14>2,8,9,10,11,13>3,15>5,6. Nessa classificação, ">" indica que a EDU anterior ao sinal é mais saliente que a EDU que o segue e, "," indica que as EDUs são igualmente importantes, pois estão no mesmo nível de profundidade na estrutura RST. A relação entre nuclearidade e saliência considerada neste modelo se apoia no fato de que os Ns de uma relação RST expressam informações mais relevantes que as que constam nos Ss e, devido a isso, são considerados mais salientes pelo conteúdo textual expresso.

Uma vez obtida a ordem de importância das EDUs, os sumários são gerados a partir da inclusão das EDUs na ordem em que aparecem na classificação de saliência até que se atinja a TC determinada.

5 O VeinSum – DETALHAMENTO E CRÍTICA DO SISTEMA

Ao receber como entrada uma estrutura RST do texto-fonte a sumarizar, o VeinSum tem por objetivo reconhecer, dessa estrutura, quais as unidades relevantes para compor o sumário e organizá-las textualmente. Além disso, seu foco é a produção de sumários que não apresentem problemas de clareza referencial. Para determinar quais segmentos serão incluídos, o sistema segue o processo ilustrado na Figura 10, adaptada de (Carbonel 2007, p.117). Seu funcionamento é reproduzido abaixo para o texto-fonte CIENCIA_2001_6423, ilustrado por TF2 no capítulo anterior.

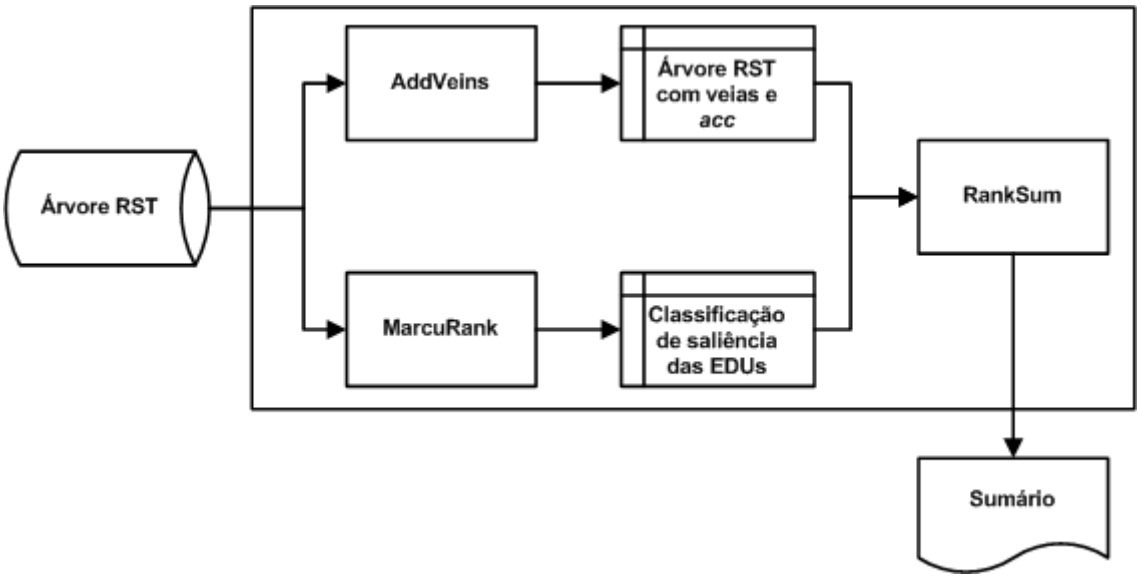


Figura 10. Arquitetura do VeinSum

A partir da árvore RST desse texto, o módulo MarcuRank determina a ordem de saliência das EDUs na árvore e o módulo AddVeins calcula as veias e o *acc* de cada uma delas. A partir desses dois resultados, o módulo RankSum é responsável pela determinação das EDUs prioritárias para compor o sumário. Para o texto em foco, o sumário produzido para a TC de 30% é ilustrado por Sum2. Em todos os sumários ilustrados nesta dissertação convencionou-se que a numeração das EDUs em negrito indica que elas são inseridas devido à classificação de saliência e a numeração normal, que elas são inseridas devido a outras regras de inserção.

[1] O mecanismo que faz as pessoas sentirem falta de ar em regiões montanhosas ou depois de uma corrida está fortemente ligado a gases da família do óxido nítrico, [11] Segundo Benjamin Gaston, um dos cientistas envolvidos na pesquisa, **[12]** a descoberta pode abrir alternativas para o tratamento de disfunções respiratórias.

Sum2. Sumário do texto CIENCIA_2001_6423

As EDUs 1 e 12 foram incluídas por terem sido apontadas como salientes pelo algoritmo de Marcu. Ao incluir a EDU 12, verificou-se que a EDU 11 é ligada a ela por uma relação de ATTRIBUTION e como é determinação do sistema que o S de uma relação desse tipo seja forçosamente incluído ao incluir o seu N, a EDU 11 passa a compor o sumário também. Essa restrição, originalmente estabelecida no VeinSum, visa a garantir que não haja perda da autoria de algum comentário incluído no sumário. No exemplo da Figura 11, o satélite "Segundo Benjamim Gaston, um dos envolvidos na pesquisa" seria incluído de qualquer forma, mesmo que não constasse do *acc*(12)={1,11,12}. Nota-se também que a ordem em que as EDUs aparecem no sumário é a mesma do texto-fonte.

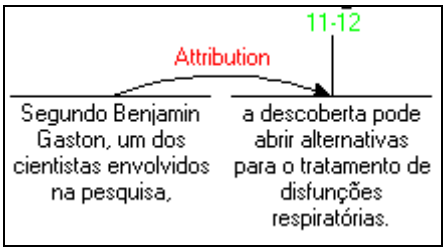


Figura 11. Estrutura RST que compreende N e S de uma relação de ATTRIBUTION[9]

Como todas as EDUs do *acc*(12) já constam no sumário, esse processo deve se repetir recursivamente até que o sistema atinja a TC determinada. Caso a inclusão da EDU 12, juntamente com seu *acc*, ultrapassasse essa taxa, haveria um descarte dessa EDU e a próxima da classificação de saliência seria testada. Os passos do VeinSum para sua inclusão de EDUs estão listados a seguir:

[9] Constuída manualmente por especialista utilizando a ferramenta de auxílio visual RSTTool - (O'Donnell, 1997), a qual fornece um suporte gráfico ao analista para a construção manual e manipulação de árvores retóricas de textos.

1) Obedecendo a classificação de saliência, elege-se a primeira EDU como candidata a inclusão;
2) Recupera seu *acc*;
3) Inclui, além da EDU saliente, seu *acc*;
4) Inclui também os *accs* das EDUs provenientes do *acc* da EDU em foco;
5) Calcula o tamanho do sumário até esse momento do processo. Se ele não viola a TC, busca a próxima EDU saliente como candidata à inclusão. Caso contrário, descarta a EDU da vez (e consequentemente seu *acc*) e elege a próxima EDU na classificação de saliência.

Além da saída principal do sistema, o sumário final, o VeinSum produz também arquivos com anotações de todo o processo de raciocínio durante a sumarização. As anotações mais importantes para a crítica ao desempenho do sistema são as que dependem dos modelos teóricos já descritos no capítulo 4. Mais particularmente, são as relativas à classificação de saliência, elaborada com base no Modelo de Saliência e proveniente do Módulo MarcuRank, e à definição dos *accs* de cada EDU de um texto-fonte, fundamentada na Teoria das Veias e proveniente do Módulo AddVeins do VeinSum (vide Figura 10 para referência). Além disso, o sistema também gera alguns resultados que refletem o conteúdo do sumário, os quais, nesta dissertação, são chamados de resultados intermediários e indicados por ResInterm.

```
ranking="[1>12>4,7,14>2,8,9,10,11,13>3,15>5>6]"
spine="[1,12]"    edus="[1,11,12]"      ignored="[]"
acc(1)={1}[10]    acc(12)={1,11,12}     origlen= "151"    maxlen= "45"
len= "55"         maxrate= "30.00"      rate= "36.42"     delta= "17.64"
```

ResInterm1. Resultados Intermediários do sumário CIENCIA_2001_6423[11]

Essa anotação fornece conjuntos de informações que indicam parte do processamento feito internamente. Observa-se em ResInterm1, que no conjunto ranking encontra-se a

[10] Os *accs* constantes dessa figura são somente os das EDUs que compõem o conjunto *spine*, porém como dito anteriormente, todas as EDU do texto-fonte têm seu *acc* determinado.

[11] Esses resultados intermediários são sempre gerados pelo VeinSum. A proposta de heurísitcas desta dissertação não gera tais resultados.

classificação de saliência para todas as EDUs do texto-fonte. Em spine, encontram-se as EDUS salientes incluídas no sumário. No conjunto edus, constam as EDUs de spine, além das que foram incluídas por serem provenientes de outras fontes que não a classificação de saliência, ou seja, provenientes dos *accs* das EDUs de spine ou ainda por serem satélites de uma relação de ATTRIBUTION. Em outras palavras, spine pode ser igual ou um subconjunto do conjunto edus. Em ignored, são agregadas as EDUs ignoradas caso a sua inclusão e, consequentemente, a inclusão de seu *acc*, corrompam a TC. Para o sumário do texto CIENCIA_2001_6423 não houve necessidade de nenhum descarte e, portanto, o conjunto ignored está vazio. Origlen indica o tamanho do texto-fonte, baseado no número de *tokens* e, portanto, o número efetivo de palavras pode ser menor. Maxlen indica o tamanho que o sumário deve ter, levando em conta a TC e len indica o tamanho do sumário gerado pelo sistema. Maxrate indica a TC determinada pelo usuário, rate indica a TC efetivamente alcançada e delta a variação da TC com relação à maxrate.

5.1 Tipos de problemas apresentados pelos sumários do VeinSum

Três tipos de problemas foram identificados nos sumários gerados pelo VeinSum: i) a ausência de segmentos antecedentes para segmentos anafóricos presentes no sumário; ii) o rebaixamento da saliência das unidades textuais do sumário, a qual ocorre quando a classificação apontada pelo Modelo de Saliência é desobedecida; iii) o corrompimento da taxa de compressão, critério mandatório na SA.

5.1.1 Ausência da EDU com antecedente mais completo

Durante a etapa inicial do trabalho, verificou-se a existência de quebras de CCRs. Entende-se por quebra quando há a inclusão de uma expressão anafórica no sumário e, pelo contexto, não é possível recuperar o seu antecedente. Essas comprometem a qualidade do texto produzido, pois acarretam dificuldades de entendimento da mensagem. Os tipos de quebras encontrados nos sumários produzidos pelo VeinSum encontram-se na Tabela 3 a seguir.

Tabela 3. Tipos de quebras de CCRs nos sumários produzidos pelo VeinSum

Pronomes Pessoais	**Ele** credita à sua colega Hannah Faye a ideia de testar de forma visual um dado já verificado verbalmente.
Pronomes Possessivos	Mas **sua** função pode ser definida de um jeito bem simples: uma balança para pesar proteínas.
Pronomes Demonstrativos	Pesquisadores da Fiocruz (Fundação Oswaldo Cruz) identificaram duas moléculas no sangue dos gambás que têm **essa** função antiofídica e esperam utilizá-las, não apenas [...].
Descrições definidas	**O instrumento** tem um nome indecifrável: espectômetro de massa de ionização por dessorção a laser com auxílio de matriz.
Entidades nomeadas (parciais)	"Para ter um genoma desse tamanho, ele precisa de grande fidelidade de replicação", afirma **Zanotto.**

Como se pode observar pelos casos listados acima, os sumários apresentam quebras de diversas naturezas, como por exemplo: de pronomes pessoais, possessivos, demonstrativos, de DDs ou de entidades nomeadas incompletas.

Após a detecção dos casos de quebras, conduziu-se uma investigação sobre a razão de tais acontecerem nos sumários e verificou-se que, na maioria dos casos, o real antecedente da expressão anafórica não está contido no *acc* da EDU que a contém.

Observa-se, em Sum3, que a anáfora "Nobre", em negrito, não tem seu antecedente "Carlos Nobre" explícito no texto e isso se deve ao fato de o *acc*(17)={1,3,17} não conter a EDU na qual está o antecedente.

> [1] O Instituto Nacional de Pesquisas Espaciais [3] prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos. [5] Esse é o pior panorama climático previsto pelo instituto. [17] **Nobre** disse [18] 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.

Sum3. Sumário parcial do texto CIENCIA_2000_17082

Isso contraria a premissa (iii**)**, de que o antecedente mais completo está sempre nos *accs* de EDUs que contêm segmentos textuais anafóricos. Porém, ao analisar algumas estruturas RST, a fim de determinar se esse problema se devia a erros no algoritmo da VT ou a árvores RST mal estruturadas, algums problemas sérios de estruturação foram detectados nas árvores, o que sugere que essa anotação deve ser revista e a comprovação da premissa (i) também não foi possível. Alguns problemas de estruturação RST encontram-se no Apêndice E,

para fins de ilustração. Caso depois de as árvores revistas e consideradas consistentes, o erro continue a ocorrer, pode-se, então, atribuir a falha ao algoritmo da VT. Carbonel (2007) reporta um estudo com precisão de 82% da VT para 12 textos do *Corpus* Summ-it, porém como a quantidade de dados é restrita, esse número pode não representar a real precisão do algoritmo em questão para textos em português.

5.1.2 Rebaixamento da classificação da saliência em prol da taxa de compressão

A partir da comparação do sumário ilustrado em Sum4, a seguir, com seus respectivos dados intermediários em ResInterm2, observa-se que as EDUs grifadas em ranking foram descartadas no momento da construção do sumário já que, se fossem incluídas, por serem longas demais, corromperiam a TC.

[1] 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, **[2]** que a comunidade científica internacional pode alterar a Declaração de Helsinque, sobre ética em pesquisas com humanos. **[3]** Em estudos no Terceiro Mundo, os cientistas se desobrigariam de fornecer aos doentes o melhor tratamento médico conhecido. **[4]** A proposta**,** [5] a ser discutida, **[6]** dá aos pesquisados o direito de receber terapia dada pelo governo de seu país **[7] -** que pode ser nenhuma.

Sum4. Sumário do texto CIENCIA_2000_17101

edus="[1,2,3,4,5,6,7]" spine="[3,2,1,4,6,7,5]"
ignored="[31,33,30,27,25,23,32,29,24,11,8,28,26,20,19,15,10,22,14,13,12,9,21,18,17,16]"
ranking="[3>16,17,18,21>2,9,12>1,4,6,13>7,14,22>10,15,19,20,26,28>5,8,11,24,29,32>23,25>27,30,33>31]"

ResInterm2. Resultados Intermediários do sumário CIENCIA_2000_17101

Outro exemplo desse caso é o sumário do texto CIENCIA_2003_24212, ilustrado em Sum5, que teve as EDUs grifadas em ranking descartadas também com vistas à manutenção da TC, como se observa em ResInterm3.

[1] Biotecnólogos e ambientalistas travaram uma rara aliança na semana passada **[3]** duas vacas deram cria em Iowa, EUA. **[6]** Os dois nascimentos, **[8]** marcam o início de uma nova fase para um projeto que já atraía interesse **[9]** - o Frozen Zoo. **[24]** A primeira tentativa de trazer um membro do Frozen Zoo de volta do mundo dos animais perdidos foi com um gauro, outra espécie rara de gado. **[25]** Uma gravidez acabou levando ao nascimento de um animal, em 2001. **[26]** O clone morreu dois dias depois. **[27]** Os dois bantengs produzidos em Iowa já viveram mais do que isso. **[42]** Os dois filhotes, **[44]** são cópias genéticas idênticas de um banteng macho que morreu no Parque Selvagem Animal de San Diego em 1980. **[45]** No início, a ACT preparou 30 óvulos de vaca, **[48]** A fusão, **[50]** faz com que o óvulo se comporte como se tivesse sido fertilizado, **[53]** Ele é então implantado numa vaca, **[55]** Clonar animais continua sendo uma tarefa difícil.

Sum5. Sumário do texto CIENCIA_2003_24212

```
<extract method ="rank- prune/marcurank/ranksum"
edus="[1,3,6,8,9,24,25,26,27,42,44,45,48,50,53,55]"
spine="[3,42,44,27,45,48,50,53,6,8,55,24,25,26,1,9]"
ignored="[40,29,39,37,34,23,14,41,33,22,18,38,32,16,12,10,52,51,49,36,35,21,20,
7,2,57,54,47,46,31,30,28,56,43,19,17,15,13,11,4,5]"
ranking="[3>42,44>27,45,48,50,53>5,6,8,55>4,11,13,15,17,19,24,25,26,43,56>1,
28,30,31,46,47,54,57>2,7,9,20,21,35,36,49,51,52>10,12,16,32,38>
18,22,33,41>14,23,34,37,39>29,40]" origlen="562" maxlen="168"
len="168" maxrate="30.00" rate="29.89" delta="-0.36">
```

ResInterm3. Resultados Intermediários do sumário CIENCIA_2003_24212

Para que o sumário ficasse com a TC em 29,89%, muitas EDUs foram descartadas, as quais, segundo a classificação de saliência eram muito mais relevantes que a EDU 9, última a ser incluída. Isso pode comprometer a relevância do sumário, já que as EDUs incluídas são muito menos salientes comparadas às descartadas.

5.1.3 Corrompimento da taxa de compressão

O corrompimento da TC também é comum nos sumários produzidos pelo VeinSum, apesar de esse ter como uma de suas prioridades respeitar ao máximo essa taxa. Observa-se que, para o sumário do texto CIENCIA_2000_17109, ilustrado em Sum6, o tamanho do

sumário, indicado por rate, ultrapassa em 10.57 o valor da TC, indicada por maxrate, como ilustra ResInterm4.

[1] Foi dado o primeiro passo para a diminuição das filas de espera para transplante de fígado. **[2]** Cientistas britânicos detectaram, em adultos, a produção de células hepáticas a partir de células-tronco da medula óssea. [16] Pesquisas em camundongos haviam mostrado que células-tronco da medula óssea poderiam originar células hepáticas, além das sanguíneas. [17] Para descobrir se o mesmo acontecia em seres humanos, **[18]** os pesquisadores analisaram células do fígado de mulheres que haviam sofrido um transplante de medula óssea, **[20]** A análise do DNA dessas células mostrou que elas continham o cromossomo Y, **[23]** O resultado deverá contribuir para o desenvolvimento de tecidos humanos para uso terapêutico,

Sum6. Sumário do texto CIENCIA_2000_17109

```
edus="[1,2,16,17,18,20,23]" spine="[2,23,18]" ignored="[]"
ranking="[2>23>18,20,24>10,16>5,7,9,12,14>1,3,11,15,17,22>4,19,21>6,13>8]"
origlen="281" maxlen="84" len="114" maxrate="30.00" rate="40.57"
delta="26.05">
```

ResInterm4. Resultados Intermediários do sumário CIENCIA_2000_17109

Além dos problemas encontrados nos sumários, inconsistências entre os próprios pressupostos do sistema foram também detectadas, como as descritas a seguir.

5.2 Inconsistências entre os pressupostos do sistema e os sumários gerados

Foram também identificados alguns problemas em que os sumários não correspondiam aos pressupostos do sumarizador, ou seja: (i) alguns dos sumários apresentavam inconsistências entre os próprios dados intermediários; (ii) alguns dos sumários não refletiam seus resultados intermediários e; (iii) outros corrompiam a determinação de inclusão de Ss da relação de ATTRIBUTION caso o N tivesse sido incluído. Esses tiveram que ser conferidos e, em alguns casos, refeitos. Isso gerou uma dúvida em relação à implementação dos algoritmos, pois as inconsistências podiam ter sido geradas caso os algoritmos tivessem sido mal aplicados. Porém, isso foi conferido e verificou-se que eles foram corretamente implementados, ou seja, a classificação de saliência e a determinação das veias e *accs* eram

confiáveis, porém o sumário final não correspondia aos resultados intermediários, como ilustram os exemplos a seguir.

5.2.1 Incoerências entre os próprios resultados intermediários

Observa-se uma inconsistência entre os *accs* das unidades constantes do sumário do texto CIENCIA_2000_6381, aqui reproduzido integralmente em Sum7, e os resultados intermediários apresentados em ResInterm5.

[1] Após o anúncio do sequenciamento do genoma, na semana passada, a França resiste como único país da União Europeia a não permitir patenteamento de genes. **[2]** A UE adota, desde junho de 1998, diretiva favorável ao patenteamento de genes. **[8]** A França é o único país que se recusa a aceitar a determinação europeia. **[19]** O assunto deve ser debatido durante a presidência francesa da UE, no segundo semestre.

Sum7. Sumário do texto CIENCIA_2000_6381

```
edus="[1,2,8,19]" spine="[1,19,2]" ignored="[]"
ranking="[1>19>2,8>3,5,10,11,13>7>6,9,15>4,17>12,14,18>16]" origlen="258"
maxlen="77" len="81" maxrate="30.00" rate="31.40" delta="4.44"
```

ResInterm5. Resultados Intermediários do sumário CIENCIA_2000_6381

Se os conjuntos spine e edus forem comparados, infere-se que a EDU 8 foi incluída no sumário, pois faz parte do *acc* de alguma EDU indicada por spine. Porém, isso não é verdade, como se pode observar pela análise dos *accs* a seguir: *acc*(1)={1}, *acc*(19)={1,19}, *acc*(2)={1,2}. Conclui-se, assim, que sua inclusão resulta do fato de ela ser saliente e, mais que isso, de ela ser a próxima candidata escolhida a partir da classificação de saliência para compor o sumário. Entretanto, essa conclusão leva à detecção de um problema de desempenho do VeinSum: como nenhum caso relativo ao *acc* se aplica, obrigatoriamente essa EDU deveria constar em spine.

5.2.2 Incoerências entre os dados intermediários do sistema e o sumário

Os exemplos abaixo são considerados inconsistências geradas pelo VeinSum, já que os sumários não refletem as informações dos dados intermediários em spine, edus e/ou ignored.

O sumário do texto CIENCIA_2000_17082, ilustrado em Sum8, com seus respectivos resultados intermediários, apresentados em ResInterm6, é um exemplo desse caso.

[1] O Instituto Nacional de Pesquisas Espaciais **[3]** prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos, **[5]** Esse é o pior panorama climático previsto pelo instituto, [17] Nobre disse **[18]** 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. **[19]** O Brasil emite 280 milhões de toneladas de carbono [21] na atmosfera por ano. [25] O desmatamento da Amazônia atingiu 16.926 km2 em 99, [31] 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.

Sum8. Sumário do texto CIENCIA_2000_17082

```
<extract method="rank-prune/marcurank/ranksum"
edus="[1,3,5,17,18,19,21,25,31]" spine="[1,3,18,5,19]" ignored="[]"
ranking="[1,3>18>5,19,21,25,31>4,9,11,17>6,15,23,29,34,36,38>2,7,12,16,22,24,
27,30,32>8,13,26,28,33,35,37,39>14,20>10] origlen="299" maxlen="89"
len="109" maxrate="30.00" rate="36.45" delta="17.71">
```

ResInterm6. Resultados Intermediários do sumário CIENCIA_2000_17082

Nesse caso, spine não reflete o conteúdo do sumário, pois as EDUs 21, 25 e 31 estão presentes no texto e não pertencem ao conjunto spine e nem ao *acc* de nenhuma EDU constante desse conjunto. Comparando-se os conjuntos spine e edus, sem qualquer referência ao sumário, já se encontraria uma inconsistência, pois as EDUs 21, 25 e 31 não poderiam ser provenientes da maior EDU de spine – a EDU 19, já que os *accs* somente compreendem EDUs anteriores à EDU em foco. Baseado nisso, duas interpretações são possíveis: 1) o VeinSum pode ter usado a veia da EDU(19)={1,3,18,19,21,25,31} e não o seu *acc*; 2) o conjunto spine está incompleto.

O sumário do texto CIENCIA_2000_17088, ilustrado em Sum9, também apresenta inconsistência, pois o conjunto spine, constante de ResInterm7, também deveria ser composto por EDUs diferentes, como se observa pela comparação com o sumário.

[1] Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de uma nova espécie de dinossauro no Brasil. [4] Batizado de Santanaraptor placidus, **[5]** o fóssil é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele. **[11]** Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o Santanaraptor ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex, [19] O exemplar de Santanaraptor encontrado pela equipe carioca foi desenterrado em 1991, [20] mas a montagem do fóssil só foi concluída nove anos mais tarde. [21] Tudo o que sobrou dele foram as patas e partes da cauda e da bacia, **[22]** mas os pesquisadores conseguiram estimar que o bicho fosse um filhote de 1,5 metro de altura.

Sum9. Sumário do texto CIENCIA_2000_17088

```
<extract method="rank-prune/marcurank/ranksum" edus="[1,4,5,11,19,20,21,22]"
spine="[1,5]" ignored="[]"
ranking="[1>5,11,22>2>3,10,16,23>6,12,17,20,26>4,7,13,18,19,21,25,27>8,14,24
>9,15]" origlen="357" maxlen="107" len="140" maxrate="30.00" rate="39.22"
delta="23.50">
```

ResInterm7. Resultados Intermediários do sumário CIENCIA_2000_17088

A inconsistência aqui se deve ao fato de o sumário conter até as EDUs 11 e 22 da classificação de saliência, as quais deveriam constar em spine. Não há como as EDUs 4,11,19,20,21 e 22, presentes no conjunto edus, serem provenientes dos *accs*(1) e (5).

Os resultados intermediários do texto CIENCIA_2000_17109, exibidos em ResInterm8, também apresentam o conjunto spine incompleto, como se pode observar pela comparação do sumário ilustrado em Sum10.

[1] Foi dado o primeiro passo para a diminuição das filas de espera para transplante de fígado. **[2]** Cientistas britânicos detectaram, em adultos, a produção de células hepáticas a partir de células-tronco da medula óssea. [16] Pesquisas em camundongos haviam mostrado que células-tronco da medula óssea poderiam originar células hepáticas, além das sanguíneas. [17] Para descobrir se o mesmo acontecia em seres humanos, **[18]** os pesquisadores analisaram células do fígado de mulheres que haviam sofrido um transplante de medula óssea, **[20]** A análise do DNA dessas células mostrou que elas continham o cromossomo Y, **[23]** O resultado deverá contribuir para o desenvolvimento de tecidos humanos para uso terapêutico,

Sum10. Sumário do texto CIENCIA_2000_17109

```
<extract method="rank-prune/marcurank/ranksum" edus="[1,2,16,17,18,20,23]"
spine="[2,23,18]" ignored="[]"
ranking="[2>23>18,20,24>10,16>5,7,9,12,14>1,3,11,15,17,22>4,19,21>6,13>8]"
origlen="281" maxlen="84" len="114" maxrate="30.00" rate="40.57"
delta="26.05">
```

ResInterm8. Resultados Intermediários do sumário CIENCIA_2000_17109

Verifica-se que a EDU 20 foi inserida no sumário e baseado nos *accs* das unidades previamente incluídas, que são: *acc*(2)={1,2}, *acc*(23)={2,23} e *acc*(18)={2,16,17,18}, a EDU 20 só pode ser proveniente da classificação de saliência e não dos *accs*, o que obrigaria que spine contivesse a EDU 20 também.

5.2.3 Casos de exclusão da Relação de ATTRIBUTION

Como se trata de regra básica do sistema, quando se tem duas EDUs interligadas por uma relação retórica de ATTRIBUTION, seu S deve ser automaticamente incluído toda vez que N o for. Porém foram encontrados exemplos em que somente o N da relação foi mantido e o S descartado. Isso gera a perda de legitimidade da informação constante no N e a ideia fica solta no texto.

O primeiro caso desse tipo é o do sumário do texto CIENCIA_2000_17108, ilustrado em Sum11, com seus resultados intermediários exibidos em ResInterm9.

[1] Um ser que invade corpos e domina a mente alheia, [3] não é mero personagem de ficção. **[4]** Para uma aranha da Costa Rica, essa criatura existe. **[6]** Apesar do nome **[8]** o tal invasor de corpos é só uma vespa. [9] O biólogo William Eberhard, da Universidade da Costa Rica, descobriu **[10]** que as larvas desse inseto, **[12]** provocam mudanças no comportamento da hospedeira. **[36]** "É uma descoberta e tanto", **[38]** "É a primeira vez que se vê uma interação química tão complexa entre parasita e hospedeiro", **[40]** A exploração alheia não tem limites. Nem mesmo no reino animal.

Sum11. Sumário do texto CIENCIA_2000_17108

<extract method="rank-prune/marcurank/ranksum"
edus="[1,3,4,6,8,9,10,12,36,38,40,41]" spine="[10,12,36,38,4,6,8,40,1,3,41,9]"
ignored="[27,21,11,7,35,32,31,28,25,22,19,15,14,5,2,39,37,34,33,30,29,26,24,23,20,18,17,16,13]"
ranking="[10,12>36,38>4,6,8,40>1,3,13,16,17,18,20,23,24,26,29,30,33,34,41>9,37,39>2,5,14,15,19,22,25,28,31,32,35>7,11,21,27]" origlen="340" maxlen="102"
len="102" maxrate="30.00" rate="30.00" delta="0.00">

ResInterm9. Resultados Intermediários do sumário CIENCIA_2000_17108

Nesse caso, as EDUs 36 e 38 são opiniões de alguém cujo referente não se encontra no sumário. O conteúdo da EDU 37 – "disse o psicólogo César Ades, da USP, especialista em comportamento de aranhas" – seria importante também, pois é informação de autoria, porém foi descartada indevidamente, apesar de ser regra interna do sistema que se preserve tanto o seu S quanto o seu N. A estrutura RST correspondente a essas EDUs encontra-se na Figura 12.

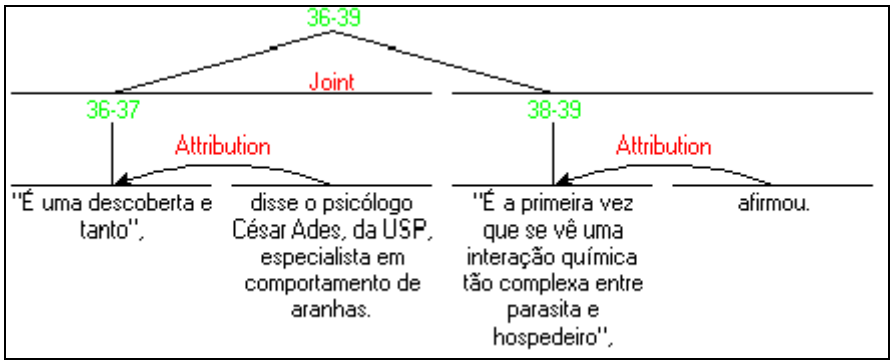


Figura 12. Estrutura RST que ilustra N e S da relação retórica de ATTRIBUTION

Outro caso de exclusão do S da relação de ATTRIBUTION é o do sumário do texto CIENCIA_2001_17109, ilustrado em Sum12, com respectivos resultados intemediários ilustrados em ResInterm10.

[1] Foi dado o primeiro passo para a diminuição das filas de espera para transplante de fígado. **[2]** Cientistas britânicos detectaram, em adultos, a produção de células hepáticas a partir de células-tronco da medula óssea. [16] Pesquisas em camundongos haviam mostrado que células-tronco da medula óssea poderiam originar células hepáticas, além das sanguíneas. [17] Para descobrir se o mesmo acontecia em seres humanos, **[18]** os pesquisadores analisaram células do fígado de mulheres que haviam sofrido um transplante de medula óssea, [20] A análise do DNA dessas células mostrou que elas continham o cromossomo Y, **[23]** O resultado deverá contribuir para o desenvolvimento de tecidos humanos para uso terapêutico,

Sum12. Sumário do texto CIENCIA_2001_17109

```
<extract method="rank-prune/marcurank/ranksum" edus="[1,2,16,17,18,20,23]"
spine="[2,23,18]" ignored="[]"
ranking="[2>23>18,20,24>10,16>5,7,9,12,14>1,3,11,15,17,22>4,19,21>6,13>8]"
origlen="281" maxlen="84" len="114" maxrate="30.00" rate="40.57"
delta="26.05">
```

ResInterm10. Resultados Intermediários do sumário CIENCIA_2001_17109

Como se pode observar na Figura 13, o conteúdo da EDU 23 é N de uma relação de ATTRIBUTION e o seu satélite "dizem seus autores" não foi incluído, o que também deixa a mensagem sem a devida legitimação.

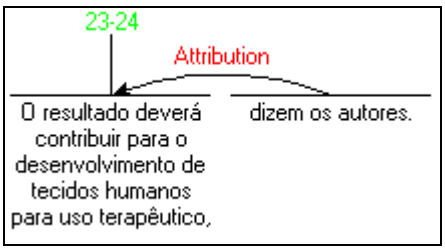


Figura 13. Estrutura RST correspondente a N e S do caso acima

Diante desses vários tipos de problemas, percebe-se que os modelos atuais do sistema são deficientes e daí a proposta de considerar a inclusão de conhecimento semântico para aprimorar o modelo de SA. Maiores detalhes sobre essa proposta encontram-se no próximo capítulo.

6 O USO DE INFORMAÇÕES SEMÂNTICAS DO PALAVRAS: EM BUSCA DO APRIMORAMENTO DA SELEÇÃO DE UNIDADES CORREFERENTES NA SUMARIZAÇÃO AUTOMÁTICA

A fim de aprimorar o VeinSum no que tange à clareza referencial de seus sumários, propõe-se o uso da etiquetagem semântica proveniente do PALAVRAS para a definição de heurísticas que objetivam melhorar o processo de seleção de EDUs para compor um sumário, mais especificamente, referindo-se à identificação de EDUs correferentes.

6.1 *Parser* PALAVRAS

O *parser* PALAVRAS é uma ferramenta de análise automática que fornece, para textos escritos em português, as informações morfossintáticas naturais de um *parsing* (lema, *part-of-speech*, gênero, número e tempo verbal) e informações semânticas. Essas informações são agregadas diretamente à representação interna de cada texto analisado sintática e semanticamente, como ilustra a Figura 14. O item lexical analisado é introduzido por word; lemma trata da sua forma canônica, dicionarizada, ou seja, os verbos são apresentados no infinitivo e substantivos e adjetivos no masculino singular. Informações de classe de palavras, conhecidas como *part-of-speech*, são indicadas por pos; morph traz informações de gênero e número, sem indica a etiqueta semântica e, finalmente, a etiqueta extra apresenta informação adicional dada por alguma das etiquetas descritas anteriormente, como no item 1 da Figura 14, no qual pos indica que o item lexical é um numeral e extra indica que o número é cardinal e não ordinal.

Para o excerto "Sete ativistas do Greenpeace foram presos ontem nos EUA ao tentar impedir o descarregamento de madeira brasileira", constante da Figura 6 (Capítulo 4), tem-se a anotação completa, na Figura 14, a seguir.

Figura 14. Exemplo de anotação completa fornecida pelo PALAVRAS

O *parser* também gera as árvores sintáticas correspondentes, como a que consta na Figura 15, referente ao excerto apresentado anteriormente. Bick (2005) reporta precisão de 99% para morfologia (classe de palavras e flexão) e 97-98% para a sintaxe, porém nada é reportado sobre os tipos de textos utilizados para essa tarefa de avaliação.

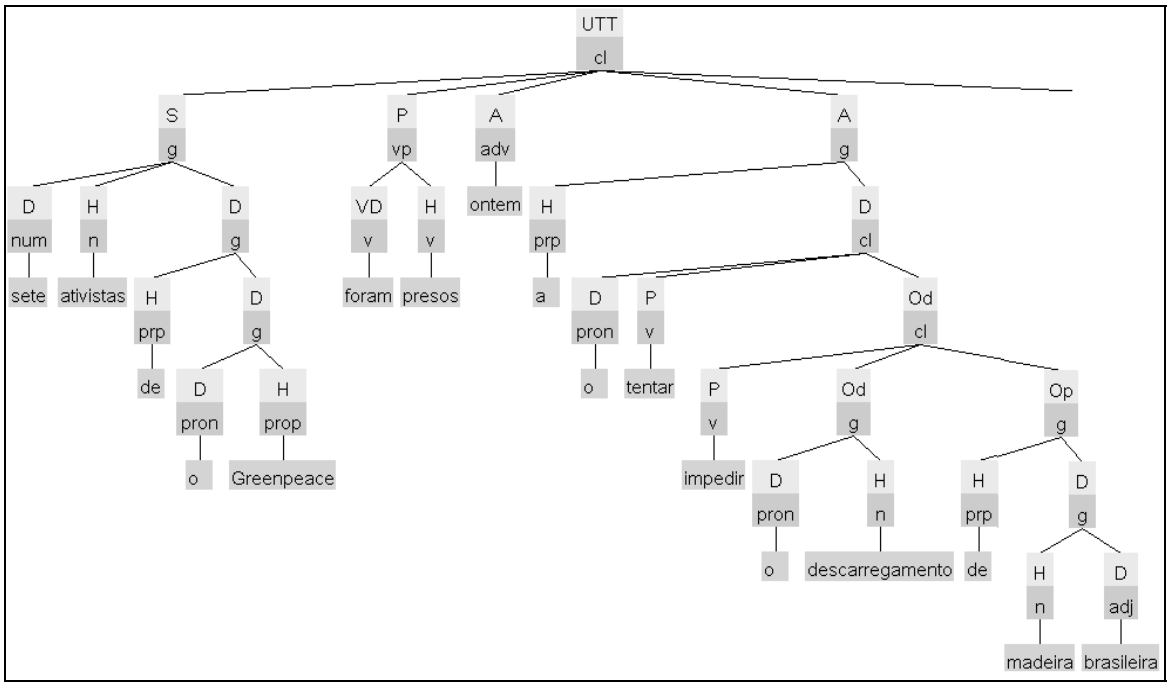


Figura 15. Estrutura sintática arbórea de parte do excerto da Figura 6

Além de fazer a anotação morfossintática, o *parser* PALAVRAS também provê etiquetagem semântica de dois tipos: a anotação de papéis semânticos, a qual engloba cerca de 40 papéis semânticos diferentes e a anotação de protótipos semânticos, que envolve cerca de 215 deles. Dentre os trabalhos que envolvem anotação semântica baseada em protótipos, tem-se o de Collovini *et al.* (2008), que consiste em classificar SNs como expressões textuais antecedentes ou anafóricas (compreendendo anáforas indiretas, diretas e associativas) a partir da anotação semântica de itens lexicais e o de Souza *et al*. (2008) que, além da anotação semântica, utiliza características sintáticas, a fim de extrair CCRs automaticamente.

Para a anotação de um item lexical considerando a etiquetagem baseada em protótipos semânticos, não se consideram modelos clássicos de semântica lexical, nos quais se buscam significados através de definições dicionarizadas ou por uma classificação ontológica, mas sim, combinações de traços semânticos, os quais fornecem uma identidade ao item lexical. Essa anotação se baseia em 16 traços, os quais supostamente representam o contexto semântico de quaisquer conceitos usados na produção de uma mensagem, segundo Bick (2000). Seu modelo semântico é definido por uma árvore de decisão, na qual cada traço tem dois ramos indicando sua ausência (-) ou presença (+) para cada item lexical, como ilustra a Figura 16 (Bick, 2000). Essa árvore é composta por 15 nós internos, que representam 12 traços semânticos diferentes (já que +/- MOVING, +/- MASS e +/- PERFECTIVE ocorrem em dois

nós) e possibilitam 16 nós folha (em itálico), os quais ilustram as possíveis categorias semânticas de quaisquer itens lexicais (ou símbolos terminais) considerados nessa classificação.

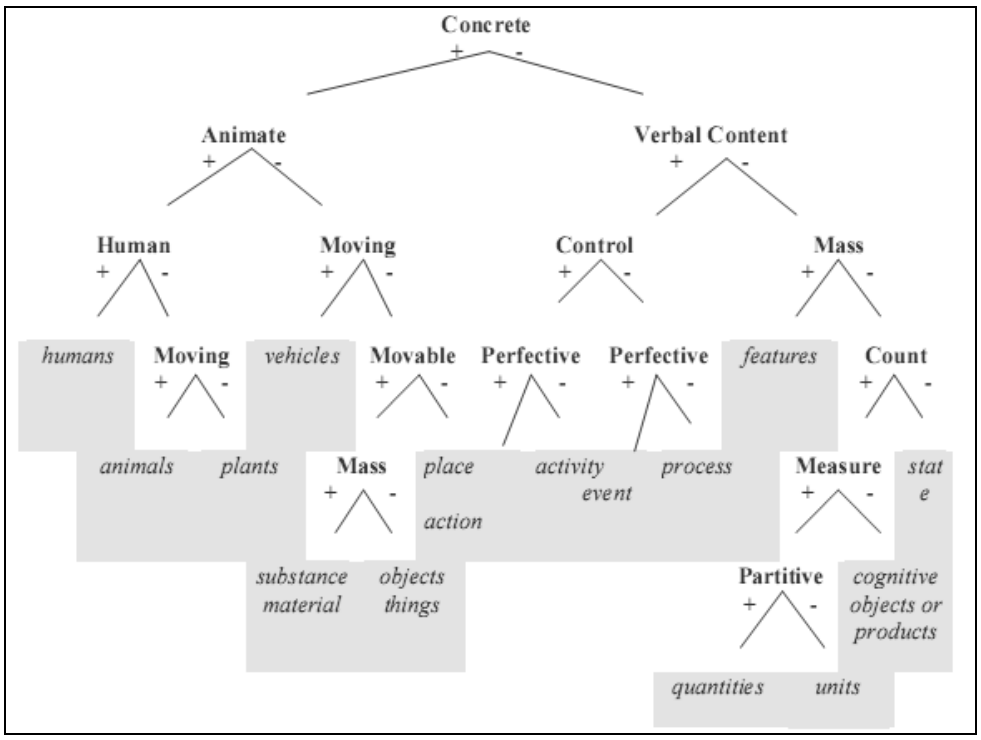


Figura 16. Árvore binária de decisão com os traços semânticos

Além dos 12 traços semânticos ilustrados na árvore de decisão, o *parser* utiliza quatro traços adicionais:

- +/- HUMAN EXPRESSION
- feature (+/- ADJECTIVAL)
- +/- LOCATION
- +/- TEMPORAL

Nesse modelo de classificação, são considerados somente os substantivos, entidades nomeadas e alguns adjetivos, para os quais é possível atribuir um valor semântico.

Os protótipos são normalmente determinados por hiperônimos de classes como, por exemplo, o protótipo animal e o protótipo flor. Esse último se enquadra no nó terminal plants, ilustrado na Figura 16, o qual funciona como um hiperônimo dos protótipos relacionados a plantas, como ilustrado na Figura 17.

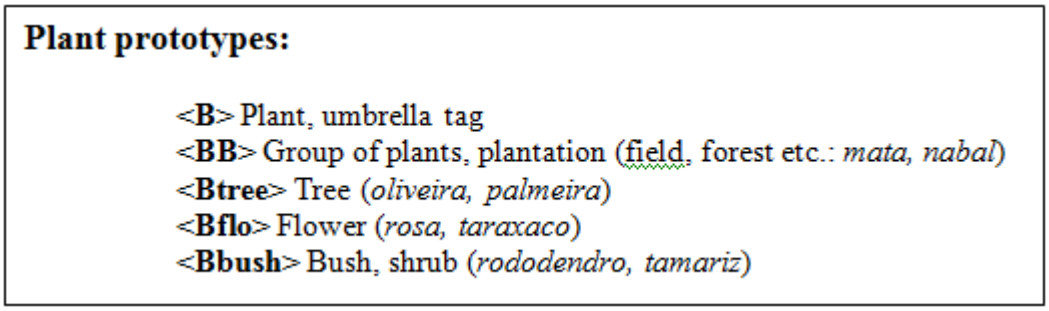


Figura 17. Protótipos relacionados a plantas

Dado um item lexical qualquer, o enquadramento não se dá através da pergunta: 'esse item lexical se refere a esse ou aquele protótipo?', mas sim através da pergunta: 'as características semânticas desse item lexical são coincidentes com as desse ou as daquele protótipo?' Assim, seus sentidos são diferenciados e a etiquetação semântica é obtida.

O item lexical animal, por exemplo, possui os traços semânticos: +Moving, -Human, +Animate e +Concrete[12]. O *parser* pode, então, etiquetá-lo, se conseguir instanciar todos os traços semânticos indicados pelo protótipo animal. Caso a ferramenta não consiga instanciar ao menos um desses traços, ele tenta enquadrar o item lexical em outro protótipo[13]. Outros traços semânticos podem ser inferidos e utilizados para desambiguação, como por exemplo: se tudo o que é etiquetado como moving, é também movable, algo que não é movable não pode ser um animal.

A identificação de itens lexicais similares é atribuida à chamada similaridade prototípica, a qual permite colocar em contexto de uso (e não em contexto definitório, usual em dicionários) a configuração semântica, sem que se necessite de coincidências absolutas de significado. Essa medida de similaridade de cada item lexical é proporcional ao número de traços semânticos que compartilham: Bick supõe que, quanto maior esse número, mais similares são os itens lexicais. Daí a possibilidade de agregar, em um único conjunto, itens lexicais similares e, em conjuntos distintos, itens dissimilares.

Assim como itens lexicais podem ser similares, protótipos também o podem. Na Tabela 4, a seguir, encontram-se agrupamentos de protótipos que compartilham os mesmos traços semânticos (indicados por *cluster*), esses representados pelas 16 letras maiúsculas no

[12] Basta seguir da folha em direção à raiz na árvore semântica da Figura 16, para recuperar os traços de animal.

[13] A listagem integral desses protótipos encontra-se no Apêndice B.

cabeçalho da tabela, com descrição correspondente. Na versão atual do PALAVRAS, existem cerca de 215 protótipos semânticos e essa tabela ilustra, portanto, somente parte deles.

Tabela 4. Protótipos que compartilham os mesmos traços semânticos [14]

E = entities (±CONCRETE) V = ±VERBAL
C = ±CONTROL P = ±PERFECTIVE
I = ±MOVING S = ±MEASURING
J = ±MOVABLE D = ±PARTITIVE
A = ±ANIMATE (living) X = ±HUMAN-EXPRESSION (allowing human modifier-ADJ)
H = ±HUMAN ENTITY F = feature (±ADJECTIVAL)
M = ±MASS L = ±LOCATION
N = number (±COUNTABLE) T = ±TEMPORAL

E	C	I	J	A	H	M	N	V	P	S	D	X	F	L	T	Cluster
+		'+	+	'+	'+	.	'+			.		+	.	.	.	H, Hprof, Hnat, Hmyth, Hfam, Htit, i, Hbio, Hsick, Hattr, *hum
+		'+	+	'+	'+	.	.			.		+	.	.	.	HH, Hhparty, *party, *media
+			.	.	'+	.	.			.		+	.	'+	.	inst, *inst
+			.	.	'+	.	'+			.		+	.	'+	.	Lciv, *civ
+		'+	+	'+	.	.	'+			.		.	.	.	.	A, Azo, Aorn, Aent, Aich, Amyth, Acell, Adom
+		'+	+	'+	.	.	.			.		.	.	.	.	AA (AAzo ..),
+			.	'+	.	.	'+			.		.	.	.	.	B, Btree, Bbush, Bflo, Bveg

'+= underivable positive feature; += derivable positive feature, .= underivable negative feature.

Em vista desse panorama, considera-se a seguinte distinção:

- Um protótipo é indicado por uma etiqueta semântica do repertório de Bick, conforme sua própria definição;
- Um agrupamento de padrões semânticos consiste em um conjunto de protótipos do PALAVRAS que compartilham os mesmos traços semânticos.

Alguns traços semânticos podem ser inferidos a partir de outros – e, assim, também as etiquetas semânticas dos itens lexicais similares. Se o item lexical 'presidente' estiver sob análise, por exemplo, seu enquadramento na classificação semântica indicará que seu protótipo representativo será o **<Hprof>** (*Professional human - marinheiro*)[15] - porque seus traços semânticos, indicados pela 16-upla dos traços considerados, remetem a

[14] Tabela obtida em contato com Eckhard Bick (versão atualizada de 2001, p. 312)

[15] As definições das etiquetas foram retiradas do elenco de etiquetas disponível em http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html#semtags e suas definições foram simplificadas nesta dissertação, com o intuito de deixar o texto mais claro.

[E,C,I,J,A,H,M,N,V,P,S,D,X,F,L,T]=[+,0,'+,'+,'+,'+,.,'+,0,0,.,0,.,.,.], constante da primeira linha da tabela. O que a tabela acima indica é que vários protótipos compartilham as mesmas possibilidades de inferência de traços semânticos.

Opostamente, agrupamentos distintos indicarão itens lexicais dissimilares. Assim é que o item lexical 'X', de etiqueta semântica **<H>** (*Human, umbrella tag*)**,** seria similar ao item lexical 'Y', de etiqueta semântica **<Hprof>**; o item lexical 'XX', de etiqueta semântica **<A>** (*Animal, umbrella tag - clone, fêmea, fóssil, parasito, predador*), seria similar ao item lexical 'YY', de etiqueta semântica **<Adom>** (*Domestic animal or big mammal - terneiro, leão/leoa, cachorro*), porém 'X' e 'XX' seriam dissimilares.

De fato, se considerada a semântica *per se*, 'X' e 'XX' são dissimilares, pois seus significados não coincidem. Eventualmente itens lexicais com significados de fato bastante distintos farão parte de um mesmo agrupamento. Por exemplo, o item lexical lasanha, de etiqueta semântica **<food-h>** (*human-prepared/complex culinary food - caldo verde, lasanha*), o item lexical acetileno, de etiqueta semântica **<cm-chem>** (*chemical substance, also biological - acetileno, amônio, anilina, bilirrubina*) e o item lexical guaraná, representado por **<drink>** (*drink - cachaça, leite, guaraná*)**,** apesar de totalmente distintos semanticamente, compartilham os mesmos traços semânticos, como se pode observar na Figura 18:

+		.	'+	.	.	'+	.			.		.	.	.	.	cm, cm-liq, cm-gas, cm-chem, cm-h, cm-rem, mat, mat-cloth, mat-h, food, food-h, food-m, fruit, drink

Figura 18. Protótipos referentes a itens lexicais semanticamente muito distintos

Nesse cenário, pode-se argumentar que o reconhecimento da proximidade semântica entre itens lexicais quaisquer introduz distorções significativas em diversas situações, já que tendo em vista o reconhecimento de CCRs, o agrupamento mencionado não reflete entidades correferentes. Ou seja, apesar de compartilharem os mesmos traços semânticos, os itens lexicais mencionados no parágrafo anterior não são correferentes, pois tratam de objetos semânticos totalmente distintos e, para os fins desta proposta, o agrupamento em questão não tem utilidade.

Para tanto, decidiu-se por uma reorganização dos agrupamentos de protótipos, através da extração de padrões semânticos que indiquem entidades realmente correferentes, a qual foi realizada através de métodos de aprendizado de máquina (Seção 6.3) a partir da anotação manual de CCRs e da anotação semântica automática pós-editada do *corpus* Summ-it. Claro é que, devido a isso, os novos agrupamentos são dependentes de *corpus* e podem ser

complementados, reduzidos ou mesmo não representar entidades correferentes presentes em outros *corpora*. Esses padrões são, então, utilizados como regras de apoio à definição de heurísticas para a geração dos novos sumários.

Objetivo da proposta

Esta proposta consiste, então, na utilização desses agrupamentos para descartar do *acc* de cada EDU escolhida para compor um sumário as EDUs que nada têm a ver com o seu contexto referencial e manter somente as que possam ser correferentes. Isso significa que, para uma EDU que contém uma anáfora, deve-se recuperar a sua etiqueta e, em seguida, percorrer o *acc* dessa EDU a fim de encontrar etiquetas iguais ou similares em alguma de suas outras EDUs componentes. Nesse caso, as etiquetas similares fazem parte do mesmo agrupamento e as EDUs que não tiverem nenhuma etiqueta igual ou similar à etiqueta da possível anáfora serão descartadas.

Como não se considera qualquer processo automático de reconhecimento de uma expressão anafórica (ou de resolução anafórica), tampouco de uma expressão antecedente nas EDUs em foco, adotou-se o SN como única construção morfossintática de interesse, para unidades lexicais candidatas a expressar ambos. Expressões anafóricas introduzidas por pronomes não são consideradas neste trabalho, pois não têm semântica própria e podem se referir a qualquer expressão textual anterior. A atribuição de etiquetas semânticas a esses itens lexicais só seria possível caso um modelo de resolução anafórica fosse considerado, pois aí os pronomes herdariam as etiquetas semânticas de seus antecedentes, o que não é o caso. Para esta proposta, isso significaria lidar com componentes textuais anafóricos, porém não anotados com etiquetas semânticas pelo PALAVRAS.

Para que os novos agrupamentos fossem delineados, primeiramente, fez-se uma análise da consistência da etiquetagem semântica, porém essa se apresentou falha porque, em parte, é dependente da segmentação do texto e da própria etiquetagem sintática, ou *parsing,* tarefas prévias realizadas pelo PALAVRAS. Afora essas causas, a própria determinação da etiqueta semântica correta se mostrou inconsistente. Isso impossibilitou a comprovação da hipótese (ii), de que a semântica do PALAVRAS era suficiente para identificar unidades textuais possivelmente correferentes. Essa anotação equivocada poderia prejudicar a validade das regras e sua consequente aplicação, o que geraria resultados não confiáveis. Decidiu-se, então, por uma pós-edição manual dessa etiquetagem, como consta a seguir.

6.2 Dados sobre a pós-edição manual da etiquetagem semântica do PALAVRAS para o *Corpus* Summ-it

Para evitar que constantes atualizações do PALAVRAS prejudicassem a consistência da tarefa de revisão manual das etiquetas semânticas, adotou-se sua versão de fevereiro de 2007 e, assim, o elenco de etiquetas semânticas e suas definições permaneceu constante.

Para a pós-edição manual do *corpus* cada um dos 50 arquivos XML resultantes da anotação semântica do PALAVRAS foi aberto em uma aba de uma única planilha Excel e novas colunas foram introduzidas para registrar informações diversas sobre as correções. Vários campos das planilhas originais geradas automaticamente encontravam-se vazios, quando deveriam conter informações. Quando elas eram evidentes, os respectivos campos foram preenchidos manualmente.

Foram sete as colunas extras adicionadas a cada planilha original. A primeira foi a de correção da segmentação das unidades frasais. Logo após a coluna com as etiquetas providas pelo *parser* (coluna value), introduziu-se a coluna 'tag sugerida'. Para o seu preenchimento, duas linguistas especialistas no elenco das etiquetas semânticas buscaram a concordância de etiquetagem, a fim de garantir a qualidade da correção dos resultados automáticos.

Todos os SNs que compõem as 589 CCRs do *Corpus* Summ-it foram verificados e, quando necessário, corrigidos, porém para a geração das heurísticas somente os seus núcleos foram computados. Tomou-se essa decisão porque, na maioria dos casos, o núcleo do SN é que determina sua informação principal e, assim, seria mais fácil determinar entidades correferentes. Por esse motivo, foi introduzida a coluna 'head', para indicar quando o componente sob análise era o núcleo de um SN. Essa coluna foi preenchida manualmente por uma das especialistas. As quatro colunas restantes das tabelas de revisão compreenderam dados de síntese para fins computacionais de geração das heurísticas e foram preenchidas automaticamente.

Após a inclusão das colunas necessárias à realização da tarefa, iniciou-se o processo de pós-edição. As decisões tomadas durante esse processo se encontram na próxima seção.

6.2.1 Decisões de Pós-edição

De forma geral, qualquer correção de etiquetas atribuídas pelo PALAVRAS aos itens lexicais somente se deu quando a etiqueta apresentou desvios semânticos consideráveis.

Nesse caso, optou-se por utilizar etiquetas mais específicas sempre que possível. Decidiu-se ainda manter etiquetas genéricas originais, se apropriadas. O objetivo foi assegurar que a avaliação do sistema refletisse seu desempenho real.

Somente foram alterados os casos em que a etiqueta era claramente indevida por ser conflitante com os traços semânticos do componente lexical, ou em que a etiqueta do item lexical fosse muito específica e, portanto, não correspondente ao seu significado adequado. Nesse caso, a nova etiqueta se referiu a um conceito mais geral. Exemplos disso ocorrem para os itens lexicais 'bicho' e 'animal', ambos etiquetados com **<Azo>** (*land animal*). Essa etiqueta claramente restringe mais do que deveria a semântica de ambos os itens lexicais, pois eles abrangem também animais aquáticos. A etiqueta atribuída foi, portanto, a genérica **<A>** (*Animal, umbrella tag - clone, fêmea, fóssil, parasito, predador*), a qual, em uma ontologia apropriada, seria considerada a superclasse da etiqueta **<Azo>**.

Para interpretar alguns itens lexicais foi necessário considerar seu contexto de uso, o que implicou reconhecer, também, os casos de delimitação de entidades nomeadas. Essas são definidas por substantivos que introduzem nomes próprios de pessoas, organizações, locais, acontecimentos, coisas, obras e conceitos abstratos. Essas categorias foram estabelecidas pela Linguateca, na taxonomia indicada para o HAREM[16]. Os componentes textuais das entidades nomeadas devem receber uma única etiqueta e, por isso, elas foram analisadas como um todo (afinal, a semântica de um componente desse tipo não é a soma da semântica de suas partes).

Se só pelo item lexical ainda não se conseguisse definir a melhor etiqueta, recorreu-se ao tópico (ou assunto) do texto, para traçar seu interrelacionamento. Por exemplo, a menção anafórica 'os pesquisados' em uma certa CCR pode se referir a pessoas, animais, medicamentos ou produtos. A partir do tópico principal do texto em que está incluída (CIENCIA_2000_17101), expresso pelo segmento "a alteração da Declaração de Helsinque, na qual os cientistas não se obrigariam a fornecer aos doentes o melhor tratamento conhecido para uma doença", é possível determinar que esse SN se refere a 'os doentes' e, portanto, é etiquetado com **<H>** (*human, umbrella tag*). Vale lembrar, no entanto, que o *parser* não propõe fazer resolução anafórica e, por isso, não tem obrigação de reconhecer esses antecedentes, mas, ao não fazê-lo, produz etiquetas que podem trazer problemas à clareza referencial dos sumários.

[16] HAREM - Avaliação conjunta na área do reconhecimento de entidades mencionadas em português - http://www.linguateca.pt/HAREM

Analisou-se também o aspecto dos itens lexicais, particularmente quando indicavam eventos, ações, atividades ou processos. Nesses casos há etiquetas específicas que distinguem a valência (+/-) do traço semântico PERFECTIVE: +PERFECTIVE indica conceito pontual; -PERFECTIVE, conceito progressivo. Distinguiram-se, ainda, as valências do traço semântico CONTROL, isto é, se os conceitos apresentados eram passíveis ou não de serem controlados. As etiquetas que tratam desses casos são indicadas abaixo:

- **<activity>** (*Activity, umbrella tag +CONTROL, IMPERFECTIVE, correria, manejo);*
- **<act>** (*Action, umbrella tag +CONTROL, PERFECTIVE);*
- **<event>** (*event -CONTROL, PERFECTIVE, milagre, morte)* e;
- **<process>** (*process -CONTROL, -PERFECTIVE*[17]*, cp.* **<event>***, balcanização, convecção, estagnação).*

Nos casos de dúvida relativas ao contexto de ocorrência e às definições das etiquetas, utilizou-se a WordNet (Fellbaum, 1998).

Além dos problemas apresentados anteriormente, vale lembrar que alguns erros de etiquetação semântica são provenientes de uma etiquetação sintática já deficiente, o que certamente provoca prejuízo maior no nível semântico.

6.2.2 Problemas de segmentação

Os casos mais problemáticos de segmentação textual do PALAVRAS residiram na confusa identificação de lexias complexas e de entidades nomeadas. Várias lexias complexas foram consideradas lexias simples, ou seja, foram processadas em componentes separados. Esse padrão foi identificado para as lexias compostas de 'substantivo + adjetivo', como nos seguintes exemplos do *corpus*:

- 'vaso sanguíneo', sendo 'vaso' etiquetado com **<con>** (*container*) e 'sanguíneo' ignorado. A lexia deveria ser etiquetada com **<an>** (*anatomical noun, umbrela tag - carótida, dorso*).
- 'cadeia evolutiva', sendo 'cadeia' etiquetada com **<inst>** (*institution*), em vez de **<ax>** (*Abstract/concept, neither countable nor mass – endogamia*) e 'evolutiva' ignorada;
- 'batimento cardíaco', sendo 'batimento' etiquetado com **<act>** (*Action, umbrella tag - +CONTROL, PERFECTIVE*), em vez de **<process>** (process -CONTROL, -

[17] Entende-se' – PERFECTIVE' como 'IMPERFECTIVE', neste caso.

PERFECTIVE, cp. **<event>**, *balcanização, convecção, estagnação*) e 'cardíaco' ignorado.

Opostamente a esses casos, a ferramenta também aglutinou vários SNs como sendo uma única entidade nomeada ou desmembrou uma única entidade em vários SNs. Os seguintes exemplos demonstram esses casos:

- Pesquisadores do Museu Nacional do Rio de Janeiro

Aqui tem-se o SN 'Pesquisadores' e as entidades nomeadas "Museu Nacional" e "Rio de Janeiro". O *parser* atribui a etiqueta **<hum>** (*person name*) para todo esse trecho, pois o considera como uma única entidade. No entanto, três etiquetas distintas deveriam ter sido atribuídas: 'Pesquisadores', com etiqueta **<Hprof>** (*Professional human – marinheiro*), 'Museu Nacional', com **<org>** (*commercial or non-commercial, non-administrative, non-party organisations*) e 'Rio de Janeiro', com **<civ>** (*civitas - country, town, state, cp.* **<Lciv>**).

- Organização das Nações Unidas

O *parser* etiquetou separadamente os seguintes itens lexicais: 'Organização', com **<np-close>**, cuja definição não é encontrada no elenco de etiquetas; 'Nações', com **<HH>** (*Group of humans - organisations, teams, companies, e.g. editora*) e 'Unidas' não recebeu etiqueta semântica alguma. Caso essa entidade nomeada não fosse desmembrada, sua etiqueta deveria ser **<org>.**

Caso as entidades nomeadas de um texto não sejam reconhecidas, a proposta de identificação de elementos correferentes por suas etiquetas semânticas se torna inviável. Esses problemas também são sérios em outras áreas de aplicação, sendo também de grande importância para a interpretação humana ou automática (Amâncio, 2009).

No total, foram identificados 104 casos problemáticos de segmentação, os quais englobam tanto a identificação de lexias complexas quanto a de entidades nomeadas. Considerando que esses componentes abrangem, em média, dois substantivos, houve 7% de casos problemáticos no *corpus* todo. Esse índice é certamente dependente do *corpus* em foco.

6.2.3 Problemas de etiquetação

Um problema significativo reside no reconhecimento de nomes de países, que devem ser etiquetados com **<Lciv>** (*Civitas, town, country, county, cidade, país*), mas muitos recebem

a etiqueta **<inst>** (*institution*). Uma possível solução para esses casos seria a utilização de um dicionário onomástico, o que não parece ser considerado no PALAVRAS.

Também nomes científicos, muito presentes nos textos do *corpus* em uso, quase sempre são etiquetados ou segmentados erroneamente. '*Tyrannossaurus rex*', p.ex., é etiquetado com **<inst>**, quando deveria receber a etiqueta **<meta>** (*meta noun - tipo, espécie*)**. <inst>** parece ser uma etiqueta usada como padrão nos casos em que não é possível encontrar uma etiqueta adequada.

Outros problemas de etiquetação seguem:

- 'células-tronco', quando inicia a oração, recebe etiqueta **<Acell>** (*Cell-animal - bacteria, blood cells: linfócito*); quando ocorre intraoracionalmente, recebe etiqueta **<HH>** (*Group of humans - organisations, teams, companies, e.g. editora*).

Independentemente da estranheza que possa causar essa variação de etiquetagem, o conflito entre **<Acell>** e **<HH>** para esse item lexical é evidente, pois a semântica de **<HH>** não é sequer aproximada à de "células-tronco", na definição original do PALAVRAS.

- 'cebola', etiquetada como **<fruit>** (*fruit, berries, nuts: maçã, morango, avelã, melancia*), em vez de **<Bveg>** (*vegetable, espargo, funcho*).

Não existe a acepção de fruta para 'cebola'. Considerou-se, aqui, um erro de etiquetagem.

- 'nave', etiquetado com **<Lwater>** (*Water place - river, lake, sea: fonte, foz, lagoa*), em vez de **<Vair>** (*Air vehicle - plane: hidroplano, jatinho*).

O mesmo caso ocorre aqui. Embora 'nave' possa se relacionar a água (acepção dicionarizada de navio), jamais o será pela ideia de localização, indicado pelo "**L**" de **<Lwater>**.

- 'cachorro' é etiquetado com **<Azo>** (*Land-animal - raposa*) e 'cão' e 'canídeos' com **<Adom>** (*Domestic animal or big mammal - terneiro, leão/leoa, cachorro*). Caso análogo ocorre com 'CO2', que recebe etiqueta **<cm-chem>** (*chemical substance, also biological - acetileno, amônio, anilina, bilirrubina*) e 'gás carbônico', que é etiquetado com **<mat>** (*material - argila, bronze, granito*, cf. **<cm>**).

Apesar de cada um desses casos indicarem itens lexicais sinônimos, as etiquetas diferentes sugerem que não há tratamento de sinonímia no PALAVRAS, o que compromete o modelo de busca de itens correferentes.

- ‘animal’ recebe etiqueta **<Azo>** (*Land-animal - raposa*)**,** o que o torna muito específico, já que nem todo animal é terrestre e deveria receber etiqueta **<A>** (*Animal, umbrella tag - clone, fêmea, fóssil, parasito, predador*).

Não há nada no *corpus* que justifique a atribuição dessa etiqueta específica a ‘animal’. Portanto, ela foi considerada um erro do *parser*.

- O item lexical ‘atmosfera’ recebe etiquetas diferentes dependendo da palavra que o segue, como em: ‘atmosfera da Terra’ e ‘atmosfera terrestre’, com etiquetas **<Ltop>** (*Geographical, natural place - promontório, pântano*) e **<sit>** (*psychological situation or physical state of affairs - reclusão, arruaça, ilegalidade, more complex and more "locative" than* **<state>** *and* **<state-h>**) respectivamente.

Há dois problemas aqui: **<sit>** não se aplica a esse item lexical e a coocorrência de ‘atmosfera’ ora com ‘da Terra’, ora com ‘terrestre’ indica sinonímia. Portanto, ‘atmosfera’ não deveria receber etiquetas diferentes, mesmo que **<sit>** tivesse alguma relação semântica com esse item lexical.

- ‘células’ recebe etiqueta **<Lh>** (*Functional place, human built or human-used - aeroporto, anfiteatro*); em vez de **<Acell>** (*Cell-animal - bacteria, blood cells: linfócito*)**.**

Não existe a acepção de lugar para ‘células’. Portanto, essa etiquetagem consiste em erro.

- ‘macho’ e ‘fêmea’ recebem, respectivamente, as etiquetas **<part-build>** (*structural part of building or vehicle - balustrada, porta, estai*) e **<A>** (*Animal, umbrella tag - clone, fêmea, fóssil, parasito, predador*).

No contexto de ocorrência, ambos os itens referem-se a **<A>** e deveriam, portanto, receber essa etiqueta.

6.2.4 Problemas de desambiguação

O processamento semântico do PALAVRAS visa à atribuição de uma etiqueta semântica que *aproximadamente* indique o significado do item lexical em foco. Para a semântica adequada, vários fatores entram em perspectiva, sendo dos mais significativos o contexto de ocorrência do item lexical e os aspectos culturais que determinam seu sentido adequado. Se o modelo semântico do *parser* pretende apontar as etiquetas semânticas aproximadas para componentes textuais, ele deveria prover mecanismos para tratar esses

fenômenos. Embora os diferenciais de tradução pretendam dar conta disso, as aproximações ainda são por demais imprecisas.

O *parser* tampouco é capaz de desambiguar itens lexicais que apresentam mais de um significado, e, em geral, determina acepções alternativas que não se aplicam, como nos casos a seguir:

- 'organismo', etiquetado com **<inst>** (*institution, functional structure +PLACE, +HUM, auto-escola, bolsa, cinemateca*), em vez de **<cc>** (*councrete countable*);
- 'clone', com etiqueta **<H>,** a qual somente se refere a clones humanos. No entanto, os contextos de ocorrência desse item no *corpus* mostram que esse termo se aplica a clones de animais e, assim, a etiqueta utilizada deveria ser a mais genérica **<A>**.

Já a desambiguação de itens lexicais em um único SN seria mais factível, devido ao contexto de ocorrência mais limitado. Porém, a forma como contextos possam ser considerados não leva a bons resultados, como ilustram os exemplos a seguir:

- 'as patas e bacia do animal', em que 'bacia' recebe etiqueta **<con>** (*container*), quando deveria receber **<anmov>** (*Movable anatomy - arm, leg, braço, bíceps, cotovelo*)*;*
- 'a física nuclear Eva Maria', em que 'física' recebe etiqueta **<domain>** (*subject matter, profession, cf.* **<genre>***, anatomia, citricultura, datilografia*), quando deveria ser **<Hprof>** (*Professional human - marinheiro*)*;*
- 'populações de pinguins', em que 'populações' recebe etiqueta **<HH>** (*Group of humans - organisations, teams, companies, e.g. editora*), em vez de **<AA>** (*Group of animals - cardume, enxame, passarada, ninhada*)*;*
- 'esqueleto do navio', em que 'esqueleto' recebe etiqueta **<Hmyth>** (*Humanoid mythical - gods, fairy tale humanoids, curupira, duende*), em vez de **<part-build>** (structural part of building or vehicle - *balustrada, porta, estai*).

Os casos que envolvem aspectos culturais demandam tratamento automático mais complexo. Em geral, essa é a principal razão de eles não serem considerados nos sistemas de PLN. Um exemplo do PALAVRAS está na etiquetagem de 'filhote' com **<H>** (*Human, umbrella tag*) quando o sentido de animal – etiqueta **<A>** – indicado pelo contexto é ignorado.

6.2.5 Crítica de desempenho do PALAVRAS

Dentre as principais dificuldades encontradas no processo de correção das etiquetas semânticas estão: i) a atribuição de etiquetas para itens lexicais de domínios específicos do

conhecimento; ii) a inadequação das definições das etiquetas e de seus exemplos, presentes no PALAVRAS; iii) o reconhecimento de etiquetas muito genéricas, muito específicas ou ainda muito abstratas; iv) a dificuldade de adequação de um item lexical a uma única etiqueta, já que muitos deles podem ser etiquetados de várias formas.

O caso (i) foi particularmente complicado, pois, apesar de o *corpus* ser de domínio geral, há textos de assuntos muito particulares para algumas áreas da ciência. Para esses, o conhecimento especialista foi crucial e as linguistas precisaram recorrer a especialistas das áreas em foco, para determinar as etiquetas que melhor refletissem a natureza dos itens lexicais.

O caso (ii) levou a uma grande dificuldade para a análise semântica, pois nem os exemplos foram suficientes para deixar claras muitas das definições. Etiquetas diferentes destinam-se a designar objetos semânticos diferentes, porém, quando se analisam os exemplos que acompanham suas definições, elas não parecem se diferenciar em nenhum aspecto. Este é o caso de **<cc-r>** (*read object - carteira, cupom, bilhete, carta, cf.* ***<sem-r>***) e **<sem-r>** (*read-work - biografia, dissertação, e-mail, ficha cadastral*), que indicam, respectivamente, uma descrição de um objeto de leitura e de um trabalho de leitura. Essas definições sugerem que o que se pretende distinguir é o modo de produção das obras escritas: **<cc-r>** seria relativa àquelas de produção simples, enquanto **<sem-r>**, às de produção complexa. Neste caso, 'e-mail' e 'ficha cadastral', por requerer produção simples, não deveriam ser exemplos de **<sem-r>.**

Há ainda etiquetas cuja definição se aplica a objetos semanticamente díspares, como **<Adom>** (*Domestic animal or big mammal - terneiro, leão/leoa, cachorro*), que, contraditoriamente, trata tanto de animais domésticos quanto de grandes mamíferos. Seria mais conveniente que essa disparidade fosse resolvida com etiquetas mais específicas, que diferenciassem animais domésticos e pequenos mamíferos de animais selvagens ou de grandes mamíferos.

Exemplos do caso (iii) são as etiquetas que, de tão específicas, têm pouca utilidade. Esse é o caso de **<anich>** (*Fish anatomy - few: brânquias, siba*) e **<cc-board>** (*flat long object - few: board, plank, lousa, tabla*), reconhecidas pelo próprio autor da ferramenta (pela palavra "few" em suas definições) como raramente aplicadas aos itens lexicais de qualquer dos *corpora* investigados.

Certamente seria esperado que tão grande elenco de etiquetas não se aplicasse integralmente a qualquer *corpus*. Entretanto, como esse grau de especificidade torna difícil a

análise de casos, questionou-se a discrepância entre alguns subconjuntos tão específicos como esses e os demais conjuntos de etiquetas. Caso similar ocorreu com as etiquetas de definições muito abstratas, como **<ac-cat>** (*Category Word - latinismo, número atômico*), corroborando o fato de que as especificações providas para o uso desse elenco não são significativamente esclarecedoras.

O fato de algumas etiquetas serem ontologicamente relacionadas dificultou o processo de revisão dos resultados automáticos, já que muitos itens lexicais podiam ser enquadrados em mais de uma etiqueta (caso (iv)). Isso ocorre, p.ex., com **<fruit>** (*fruit, berry, nut - still mostly marked as* **<food-c>***, abricote, amora, avelã, cebola*) e **<food-c>** (*countable food - few: ovo, dente de alho, most are* **<fruit>** *or* **<food-c-h>** *culinary countable food - biscoito, enchido, panetone, pastel*). Certamente, as duas etiquetas são apropriadas para alguns itens lexicais, porém optou-se por utilizar a etiqueta mais específica neste caso.

Além dos casos acima, ocorrências menos significativas, mas não desprezíveis do ponto de vista da proposta semântica do PALAVRAS, foram elencadas. Verificou-se, dentre elas, que o elenco das 215 etiquetas não foi suficiente para descrever alguns itens lexicais comuns. 'vírus', por exemplo, é etiquetado inadequadamente com **<Acell>** (*Cell-animal - bacteria, blood cells: linfócito*), pois não é um animal celular, mas sim "uma partícula proteica que infecta organismos vivos"[18]. A etiqueta mais próxima a ser atribuída a esse item lexical seria **<cc>** (*concrete countable*), porém, por ser muito genérica, ficou difícil determinar, pelo contexto, sua aplicabilidade. Decidiu-se, assim, manter **<Acell>.** Esse foi o único caso de manutenção de etiqueta quando claramente imprópria.

Outras etiquetas são classificadas por Bick como *vazias*, como **<cc-h>** (*artifact, umbrella tag - so far empty category in PALAVRAS*) e parecem se associar a casos não previstos (indicação dada pelo termo *umbrella tag*). No entanto, na ausência de etiquetas adequadas, a escolha pelas ditas *vazias* foi considerada.

Há ainda as marcadas como 'Further proposed categories', para as quais não há definições ou não há exemplos, constituindo-se, assim, em etiquetas subespecificadas. **<spice>** é um caso de ausência completa de descrição; **<top>** (*geographical location*) e **<Bveg>** (v*egetable, espargo, funcho*), de subespecificação.

[18] http://pt.wikipedia.org/wiki/Vírus (Acesso em 25 jun. 2009).

O uso da etiqueta **<meta>** (*meta noun - tipo, espécie*) também não ficou claro. A referência a tipo ou espécie sugere a possibilidade de se recorrer a uma relação ontológica. Desse modo, ela poderia ser utilizada para itens lexicais que indicam, por exemplo, classe, gênero ou raça (hiperônimos) de 'equinos' ou 'manga-largas' (hipônimos correspondentes). Decidiu-se por utilizá-la para ocorrências de ambos os tipos, já que nenhuma outra etiqueta do elenco seria apropriada para cobrir esses casos.

Os critérios relatados nesta seção foram adotados mediante a necessidade de se buscar etiquetas adequadas a cada caso. Procurou-se restringir ao máximo as alterações de anotações originais do *parser*. Ressalta-se ainda que todas as etiquetas constantes do elenco foram utilizadas na pós-edição, quando se julgava que sua definição era adequada para enquadrar os itens lexicais.

Na seção seguinte, encontram-se os dados numéricos da pós-edição manual da etiquetagem gerada automaticamente.

6.2.6 Síntese numérica dos casos

O texto com menor porcentagem de correção apresentou aproximadamente 3% de problemas de etiquetação e o com maior índice, aproximadamente 67%. A porcentagem média de correção do *corpus* foi de 41%. Além disso, observou-se também que as CCRs referentes a pessoas, as quais, em geral, incluem nomes próprios e profissões, são as que apresentam maior índice de acerto.

Os textos menos problemáticos para a etiquetagem semântica são os de domínios mais genéricos e os que apresentam porcentagem maior de correção são os de domínios mais específicos, como ilustra a Tabela 5.

Tabela 5. Quadro da correção da etiquetagem semântica para o *corpus* Summ-it

	Texto-fonte	# total de subs	# subs etiquetados corretamente	# subs corrigidos	# erros de segmentação	% de correção
1	CIENCIA_2005_6507	24	8	16	2	**66,67**
2	CIENCIA_2003_6465	41	14	27	4	**65,85**
3	CIENCIA_2003_24212	106	41	65	4	**61,32**
4	CIENCIA_2001_19858	63	25	38	6	**60,32**
5	CIENCIA_2005_28752	72	29	43	2	**59,72**
6	CIENCIA_2001_6410	27	12	15	2	**55,56**
7	CIENCIA_2000_17088	62	28	34	4	**54,84**
8	CIENCIA_2001_6423	17	8	9	1	**52,94**
9	CIENCIA_2002_22029	99	47	52	1	**52,53**
10	CIENCIA_2002_6441	21	10	11	0	**52,38**
11	CIENCIA_2000_6381	60	29	31	1	**51,67**
12	CIENCIA_2000_17113	76	37	39	1	**51,32**
13	CIENCIA_2005_28764	98	48	50	0	**51,02**
14	CIENCIA_2000_17108	55	27	28	1	**50,91**
15	CIENCIA_2000_6389	31	16	15	0	**48,39**
16	CIENCIA_2004_26417	52	27	25	7	**48,08**
17	CIENCIA_2005_28755	82	44	38	2	**46,34**
18	CIENCIA_2000_17101	59	32	27	1	**45,76**
19	CIENCIA_2002_22023	60	33	27	1	**45,00**
20	CIENCIA_2005_28754	65	36	29	4	**44,62**
21	CIENCIA_2002_22015	70	39	31	3	**44,29**
22	CIENCIA_2004_6480	50	28	22	0	**44,00**
23	CIENCIA_2003_24226	84	48	36	3	**42,86**
24	CIENCIA_2005_28756	75	44	31	0	**41,33**
25	CIENCIA_2001_6414	30	18	12	1	**40,00**
26	CIENCIA_2004_26415	33	20	13	1	**39,39**
27	CIENCIA_2005_28766	107	65	42	8	**39,25**
28	CIENCIA_2002_22027	91	56	35	1	**38,46**
29	CIENCIA_2000_17082	37	23	14	1	**37,84**
30	CIENCIA_2000_17109	75	47	28	1	**37,33**
31	CIENCIA_2004_6494	30	19	11	7	**36,67**
32	CIENCIA_2005_6515	41	26	15	0	**36,59**
33	CIENCIA_2003_6472	22	14	8	0	**36,36**
34	CIENCIA_2005_28744	85	55	30	0	**35,29**
35	CIENCIA_2000_17112	54	36	18	6	**33,33**
36	CIENCIA_2001_6406	21	14	7	0	**33,33**
37	CIENCIA_2005_6514	37	25	12	0	**32,43**
38	CIENCIA_2004_26423	115	78	37	10	**32,17**
39	CIENCIA_2005_6518	45	31	14	0	**31,11**
40	CIENCIA_2004_6488	13	9	4	0	**30,77**
41	CIENCIA_2001_6416	43	30	13	1	**30,23**
42	CIENCIA_2000_6391	41	29	12	2	**29,27**
43	CIENCIA_2005_28747	42	30	12	0	**28,57**
44	CIENCIA_2000_6380	31	23	8	4	**25,81**
45	CIENCIA_2004_26425	99	76	23	1	**23,23**
46	CIENCIA_2003_24219	77	60	17	4	**22,08**
47	CIENCIA_2002_22005	62	50	12	4	**19,35**
48	CIENCIA_2002_22010	36	30	6	0	**16,67**
49	CIENCIA_2003_6457	45	39	6	2	**13,33**
50	CIENCIA_2005_28743	35	34	1	0	**2,86**
	Médias	**2796**	**1647**	**1149**	**104**	**41,09**

Essa variação grande da porcentagem de correção para textos de domínios genéricos e específicos pode se dever ao fato de o módulo de etiquetagem semântica ter sido fundamentado, em grande parte, na chamada "diferenciação de equivalentes de tradução" (Bick, 2000, p. 5)[19], evitando-se, assim, aspectos definicionais ou a adequação referencial, além da utilização de um *corpus* específico de tradução, cujas características não se coadunam com as do *Corpus* Summ-it, o que exigiria, portanto, uma retreinagem do modelo semântico do PALAVRAS.

Derivou-se, portanto, um conjunto de dados pós-editados, no que se refere ao processamento automático do PALAVRAS, e mais rico, no tocante à análise semântica especializada. Por esse motivo, além de servir como objeto de projeto de sistemas de SA, ele serve também como *corpus* de referência para várias tarefas, quer de avaliação ou validação, quer de linguística de *corpus*. Pode servir, ainda, como modelo para a revisão das decisões do próprio *parser,* caso se pretenda refiná-lo para contemplar os casos incorretos apontados na crítica aqui apresentada. Em outras palavras, ele é um repositório importante e útil para várias áreas do conhecimento.

Vale destacar que o trabalho de engenharia linguística para a correção dos dados foi grande, já que 2796 substantivos foram verificados e a porcentagem de correção dessa etiquetagem chegou aos 41%, como indicam os dados gerais desse processo de correção na Tabela 5. Maiores detalhes sobre essa tarefa encontram-se em Tomazela & Rino (2010). Bick (2007) reporta precisão de 90.5% da anotação de papéis semânticos para 2.500 palavras em Português europeu extraídas do Floresta Sintá(c)tica Treebank, porém não reporta nenhuma avaliação dos protótipos semânticos e, portanto, não é possível fazer comparação.

Na sequência da pós-edição manual, as heurísticas foram obtidas através de um modelo de aprendizado de máquina. Esse processo e a filtragem dos dados relevantes encontram-se descritos a seguir.

6.3 Heurísticas de seleção de EDUs

Como o modelo semântico em uso remete às definições das etiquetas do PALAVRAS, a engenharia de conhecimento aqui sugerida adota os pressupostos conceituais desse modelo.

[19] Tradução nossa.

Considerou-se ser possível, pela análise de *corpora* textuais, reproduzir agrupamentos de protótipos que pudessem indicar itens lexicais correferentes e, assim, tentar comprovar as hipóteses (i) e (ii) deste trabalho, de que itens lexicais correferentes apresentam etiquetas semânticas iguais ou similares e que a pós-edição da etiquetagem semântica do PALAVRAS é suficiente para identificar unidades textuais possivelmente correferentes. Para que essa identificação fosse possível, consideraram-se regras baseadas nas informações semânticas provenientes do PALAVRAS, as quais foram geradas automaticamente pelo Weka (Hall *et al.*, 2009)[20].

Com o *Corpus* Summ-it etiquetado semanticamente como entrada, variaram-se os métodos de aprendizagem para gerar um conjunto de regras de associação e um conjunto de regras de classificação. Esses conjuntos constam nos apêndices C e D, respectivamente, porém, são modificações dos gerados pelo Weka após uma análise manual. Vale ressaltar que as chamadas regras de associação, neste trabalho, são compostas apenas pelos seus item sets e não correspondem às regras finais geradas pelo Weka. As regras de associação são identificadas por "A" e as regras de classificação, por "C". Caso se considerem várias regras de mesmo tipo, seus identificadores estão numerados. Por exemplo, A1, A2, C1, C2, C3, etc.

Foram excluídos do *corpus* todos os SNs que não compunham CCRs ou que fossem pronominais. O conjunto restante de SNs consistiu de 216 instâncias de agrupamentos de etiquetas semânticas (uma etiqueta para cada N do SN em foco), as quais foram utilizadas como entrada do Weka, para o aprendizado automático das regras. Em outras palavras, cada uma das 216 instâncias passou a ser representada como uma "215-upla" em que cada um de seus termos, correspondente a uma das 215 etiquetas semânticas do elenco completo do PALAVRAS, indicava a presença ou ausência da etiqueta em questão.

Um exemplo da construção dessa 215-upla é dado para o conjunto de etiquetas semânticas **<mat>**, **<cc-board>**, **<amount>**, **<activity>** e **<cc>**, as quais, em ordem alfabética, ocupam, respectivamente, as posições 13 (**<activity>**), 22 (**<amount>**), 48 (**<cc>**), 50 (**<cc-board>**) e 146 (**<mat>**). O vetor completo pode ser expresso, por exemplo, por:

$$<no,\ldots,no,\mathbf{v_{13}},no,\ldots,no,\mathbf{v_{22}},no,\ldots,no,\mathbf{v_{48}},no,\ldots,no,\mathbf{v_{50}},no,\ldots,no,\mathbf{v_{146}},no\ldots>$$

em que cada v_i indica a posição em que a etiqueta está presente, ou seja, v_i = *yes* em todos os casos.

[20] Disponível em: http://www.cs.waikato.ac.nz/ml/weka

Para as regras de associação, uma instância de entrada para o Weka é exatamente como ilustrado acima: não se distingue nenhuma etiqueta associada à CCR em foco na 215-upla. Já para regras de classificação, das 215 etiquetas, uma delas indicará a classe à qual as demais serão relacionadas. Neste caso, o vetor é composto de uma posição extra – a da classe, que será a última posição do vetor. Na posição da etiqueta que representa a classe, dentre as 215 posições anteriores do vetor, não haverá indicação de presença da etiqueta naquela instância, ou seja, o valor, nessa posição, será igual a *no* também. Assim, suponha-se que o mesmo conjunto de etiquetas semânticas **<mat>**, **<cc-board>**, **<amount>**, **<activity>** e **<cc>** represente uma instância de etiquetas de uma CCR do *corpus* e que a classe que as representa seja **<mat>**. As posições 13 (**<activity>**), 22 (**<amount>**), 48 (**<cc>**) e 50 (**<cc-board>**) da 216-upla, agora, terão valor *yes.* As demais posições, inclusive a 146, conterão *no.* Entretanto, a última posição desse vetor conterá a etiqueta que representa a classe, ou seja**, <mat>**. O vetor completo, neste caso, será expresso, agora, por:

$$<no,\ldots,no,\mathbf{v_{13}},no,\ldots,no,\mathbf{v_{22}},no,\ldots,no,\mathbf{v_{48}},no,\ldots,no,\mathbf{v_{50}},no,\ldots,no,\mathbf{v_{146}},no\ldots,no,\boldsymbol{mat}>^{21}.$$

Para cada instância considerada, a classe indica a etiqueta semântica do que deve ser o antecedente mais completo daquela CCR. No caso, tomou-se a etiqueta do primeiro componente da cadeia, sob a suposição de que, em um texto claro, o antecedente mais completo está sempre presente e é a primeira menção ao referente.

A principal diferença entre as regras de associação e as de classificação é que um conjunto de etiquetas de uma CCR, no primeiro caso, agrega as etiquetas que, pelo *corpus*, coocorrem mais frequentemente sem, contudo, discriminar qual é a etiqueta do provável antecedente. Já as regras de classificação agregam essa discriminação.

Como exemplo geral da diferença entre os dois conjuntos gerados pelo Weka, suponha-se que em uma CCR qualquer as etiquetas dos Ns de seus SNs sejam: $etiqueta_1$, $etiqueta_2$, $etiqueta_3$, ...,$etiqueta_n$. Suponha-se também que o antecedente mais completo apresente a $etiqueta_1$, nesse conjunto de etiquetas. A regra de associação gerada pelo Weka, a partir dessa CCR, seria a regra A:

(A) $etiqueta_1$, $etiqueta_2$, $etiqueta_3$, ..., $etiqueta_n$

[21] Certamente qualquer dos v_i indicados neste vetor será igual a *yes.*

Já a regra de classificação gerada a partir da mesma CCR seria uma implicação lógica em que as premissas indicam a ocorrência de todas as etiquetas de Ns de SNs nessa CCR exceto a $etiqueta_1$, a qual será a classe indicada pelo Weka como etiqueta do antecedente mais completo. Na implicação lógica, ela figura como conclusão, ou consequência lógica, portanto:

(C) $etiqueta_2$, $etiqueta_3$, ..., $etiqueta_n$, ➔ $etiqueta_1$

A parametrização do Weka foi realizada da seguinte forma: para as regras de associação utilizou-se o método *A priori*, o qual permite definir os parâmetros de suporte e confiança, respectivamente definidos pelo número de instâncias usadas para se ter um resultado confiável e pelo impacto dos resultados na base de dados. Como a base de instâncias era muito pequena (216 casos somente), o parâmetro confiança adotado foi anulado, para que houvesse possibilidade de se gerar algum resultado útil. Para as regras de classificação foram considerados os métodos J48, ONER e PART, porém, também devido ao tamanho reduzido da base, uma análise da similaridade entre os resultados dos três métodos levou à opção pelos resultados do J48. Em ambos os casos não houve qualquer possibilidade de suporte à significância estatística desses resultados. Apesar disso, a proposta de uso desse ambiente de aprendizado foi considerada válida, como apoio para o estudo linguístico proposto nesta pesquisa.

Assim, os dois conjuntos gerados pelo Weka foram considerados para a geração das heurísticas, embora não tenham sido usados diretamente como foram gerados. Eles passaram por uma filtragem manual prévia, realizada por uma especialista no elenco das etiquetas semânticas do PALAVRAS.

Ao final do processamento dos conjuntos de itens das regras de associação realizado pelo Weka foram obtidos 281 conjuntos de *item sets*. Após a crítica pela especialista, 136 *item sets* foram excluídos do conjunto por não compreenderem etiquetas de fato correferentes ou serem pouco frequentes e, portanto, insignificantes para esta proposta. O conjunto restante conta com 145 regras de associação de *item sets* e está listado no Apêndice C.

Dentre as regras excluídas está, por exemplo, (**<cc>,<am>**). As definições das etiquetas são: **<cc>** (*Concrete countable object, umbrella tag - briquete, coágulo, normally movable things*) e **<am>** (*Abstract mass/non-countable, umbrella tag - habilidade, legalidade*). Por suas definições, essas etiquetas não podem coocorrer já que uma se refere a um conceito abstrato e a outra a um conceito concreto.

Vale ressaltar que, além dos motivos já citados, outro que leva à exclusão de regras é o fato de se considerar somente os Ns dos SNs em correferência, pois, muitas vezes, esses sozinhos não são suficientes para explicitar a relação de correferência entre os SNs do texto. Por exemplo, em uma CCR com as seguintes menções:

- O aumento da temperatura de até 5º C nas áreas mais secas da Amazônia;
- O pior panorama climático previsto pelo instituto.

Como são considerados somente os Ns dos SNs, grifados acima, é difícil fazer um mapeamento entre aumento e panorama e, consequentemente, a regra que compreende as etiquetas **<process>** e **<ac>**, correspondentes a esses itens lexicais, foi excluída do conjunto.

Já para as regras de classificação, a partir dos resultados do Weka, obteve-se 178 regras, das quais muitas ocorriam mais de uma vez, dependendo do processamento realizado, e outras tinham frequência muito baixa ou não compreendiam etiquetas de fato correferentes, como ocorrido também no conjunto de regras de associação. Esse conjunto passou pela filtragem manual e somente 37 regras foram mantidas.

Uma regra considerada no conjunto novo de classificação é a **C1**, dada pela seguinte implicação lógica:

civ = no AND
per = no AND
hprof = yes AND
hum = no: hum

Essa regra pode ser lida da seguinte forma: se uma CCR não contiver Ns de SNs etiquetados com **<civ>**, **<per>**, **<hum>**, mas contiver um deles etiquetado com **<Hprof>**, então a possível etiqueta do seu antecedente mais completo será **<hum>**. A classe, ou etiqueta do suposto antecedente mais completo, é indicada após o sinal ":", no formato resultante do Weka.

Todas as premissas que envolviam presença (as iguais a 'yes', no caso) e ausência de etiquetas (as iguais a 'no') como no exemplo acima indicaram a impossibilidade de coocorrência de ambos os grupos de etiquetas, por sua definição semanticamente díspare. Assim, elas jamais poderiam corresponder a itens lexicais correferentes. No entanto, as marcadas com 'yes' de fato apresentaram compatibilidade semântica. Por isso, decidiu-se

desconsiderar das regras toda premissa indicativa de ausência de etiqueta (as marcadas com 'no') e manter somente as de tipo 'yes' nas regras de classificação[22]. Portanto, o resultado da simplificação de (C1) é a implicação lógica 'hprof → hum'. A ideia é, então, indicar uma CCR pela associação de possíveis etiquetas semânticas correferentes.

Além das regras pouco significativas e das que não indicavam expressões de fato correferentes, foram também desconsideradas do conjunto de classificação as regras de única premissa do tipo 'etiqueta = no", como a ilustrada abaixo, pois não indicam etiquetas de anáforas especificamente, somente de antecedente.

pub = no AND
vair = yes: v

Vale notar que, embora todas as etiquetas utilizadas para os Ns dos SNs tenham sido consideradas, muitas delas não tiveram representatividade significativa no *corpus*, levando em conta o fenômeno da coocorrência e, portanto, essas etiquetas não constam de nenhuma regra A ou C.

O funcionamento do modelo que utiliza tais heurísticas é descrito na próxima seção.

6.4 Detalhamento do modelo proposto

Este novo modelo propõe que, agora, cada EDU escolhida para compor o sumário seja indicada pelas seguintes informações: sua classificação de saliência, sua ocorrência no *acc* de alguma EDU já escolhida e a satisfação de alguma heurística que envolva suas etiquetas semânticas.

A aplicação deste modelo de raciocínio parte, assim, de uma EDU saliente e leva, a cada análise de *acc*, à escolha das EDUs que apresentarem maior grau de similaridade semântica com a EDU saliente em foco. São as heurísticas de identificação de EDUs mais similares semanticamente a cada EDU saliente candidata a inclusão no sumário as responsáveis pelo novo raciocínio.

O novo modelo supõe que o sistema de SA recebe como entrada, além da árvore RST, o texto-fonte anotado previamente pelo PALAVRAS com as etiquetas semânticas (dado pelo

[22] De fato, considerando-se a lógica do conjunto de premissas, se as etiquetas marcadas com 'no' nunca são correferentes, cada premissa desse tipo será tautológica e, assim, neutra na conjunção de premissas, podendo ser retirada por simplificação.

pontilhado na Figura 19). Porém, por se tratar de um trabalho teórico e, portanto, manual, foram considerados somente os SNs que estão em relação de correferência com outros SNs do texto; informação retirada da anotação de correferência do *corpus* Summ-it. Esse recorte do cenário ideal foi necessário já que, sem o modelo heurístico implementado, as combinações entre etiquetas seriam inúmeras para se fazer mentalmente, o que tornaria os resultados muito passíveis de erros. Assim, considera-se, além do TF anotado com informações semânticas, as CCRs anotadas manualmente. Porém, vale ressaltar que, caso o modelo heurístico seja implementado, não será necessária a utilização de informações de CCRs como entrada para o módulo de sumarização.

O novo componete heurístico de decisão para a filtragem de EDUS que contém antecedentes mais prováveis é incorporado ao módulo denominado Sumarização[23].

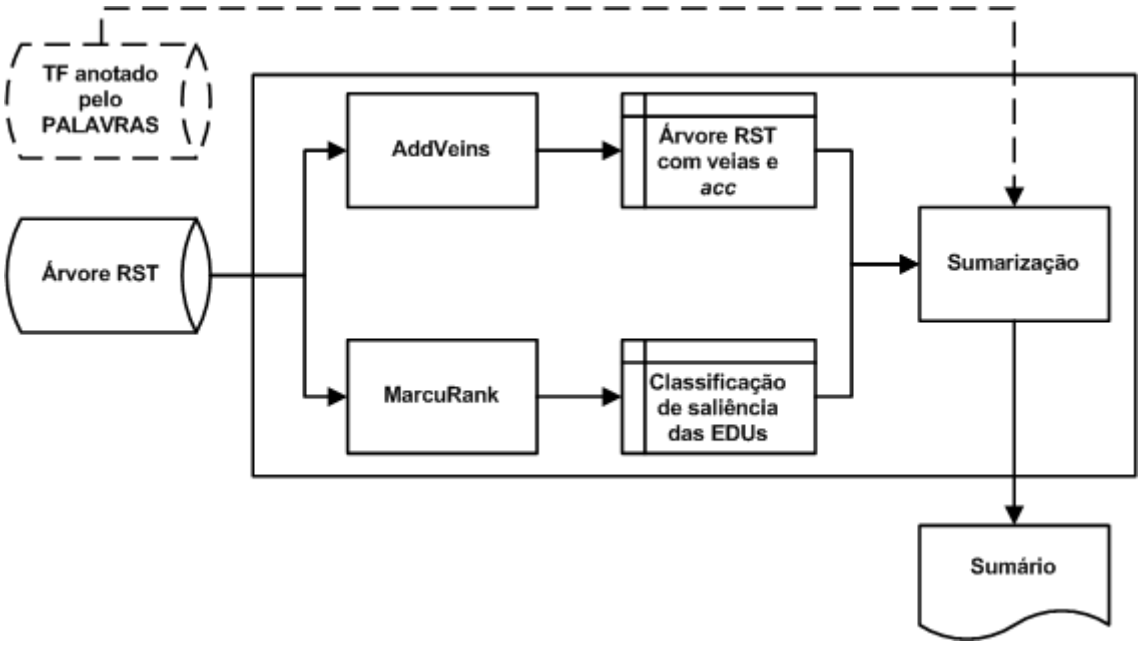


Figura 19. Arquitetura do modelo proposto

Seja a classificação de saliência original de n EDUs de certo texto-fonte dada, por exemplo, por EDU_1**,** EDU_2 **,** EDU_3 > EDU_4 **,** EDU_5 > **...>** $\mathbf{EDU_j}$ > EDU_{j+1} **,** EDU_{j+2} **,** EDU_{j+3} **>...>** EDU_n[24] e seja a EDU_j a próxima EDU escolhida para inclusão no sumário.

[23] Compare essa arquitetura com a do VeinSum, na Figura 10.

[24] Os subscritos, aqui, indicam somente a ordem de saliência e não o número da EDU, ou sua posição, no texto-fonte.

De modo geral, considera-se que o sumário em construção já tenha alguns segmentos textuais e, portanto, tenha um tamanho calculado T_0. Esse sumário será identificado aqui por *sumário anterior*. Chegado o momento de incluir a EDU_j, o procedimento a ser realizado para se obter um *sumário novo* é o seguinte:

- Se a EDU_j ou qualquer EDU de seu *acc* aparecer em uma relação de ATTRIBUTION ou SAME-UNIT com outra EDU já incluída no sumário, ela deve ser incluída incondicionalmente, para que não se perca a autoria da mensagem proferida no primeiro caso e para que a proposição inteira seja preservada, no segundo.
- Busca-se o *acc* da EDU_j, aqui denotado por *acc*(j). Como esse conjunto é composto pelas EDUs na ordem em que elas aparecem no texto (e, portanto, ordenado crescentemente com base nos índices das EDUs), a EDU_j é a última EDU do *acc*(j). Sendo esta a EDU que pode conter uma anáfora, somente as EDUs que a antecedem no *acc(j)* (isto é, com índices menores que j) poderão ser candidatas à verificação de similaridade semântica.
- Verifica-se a similaridade semântica entre as etiquetas dos núcleos dos SNs da EDU_j e as etiquetas dos núcleos dos SNs de cada EDU de seu *acc*, de acordo com a seguinte sequência de decisões:
    - Sob a hipótese de que um antecedente de uma possível anáfora ocorra próximo a ela (Mitkov, 2002) (Chaves, 2007), a primeira EDU analisada deve ser a mais próxima da EDU_j no *acc*(j).
    - A análise de cada EDU candidata recai sobre as EDUs do *acc*(j) que ainda não constam do sumário. Entretanto, se a EDU_j, já estiver no sumário, não é necessário considerar nenhuma outra EDU candidata.
    - Aplica-se alguma das heurísticas do modelo proposto. Consideram-se, primeiramente, as regras de associação, pois elas cobrem os casos de classificação, exceto em 3 casos, dos 37 desse conjunto. Vale lembrar, porém, que, na associação de etiquetas, somente sua coocorrência é considerada, perdendo-se a informação dada pela implicação lógica, que indicaria a etiqueta mais adequada a considerar. As heurísticas em foco são as que envolvem as etiquetas de algum núcleo dos SNs da EDU_j e da EDU candidata. Basta contemplar ao menos um par de etiquetas da EDU sob análise e da EDU_j para que essa seja considerada candidata a inclusão no sumário.

- Se nenhuma regra de associação se aplicar, buscar a coincidência direta de etiquetas da EDU_j e da EDU candidata.
- Se não for possível encontrar etiquetas coincidentes, usar alguma regra de classificação que se aplique ao caso, ou seja, aquela com premissa que indique a etiqueta do N da EDU_j em foco e com consequência lógica que indique a etiqueta do possível antecedente. Essa etiqueta deve ser associada a um dos componentes da EDU candidata.
- Recuperado o elenco de EDUs candidatas, as que forem consideradas semanticamente similares à EDU em foco devem ser incluídas no sumário. Porém caso a EDU_{j-1} apresentar um candidato a antecedente e a EDU_{j-2} apresentar três candidatos para as 4 anáforas da EDU_j, a EDU_{j-2}, em vez da EDU_{j-1}, deve ser a escolhida.

- Se nenhum dos casos anteriores se aplicar, manter todo o *acc* da EDU saliente no sumário, isto é, incluir todas as suas EDUs que ainda não constam no sumário.

No caso geral, se durante a análise de EDUs do *acc*(j) a EDU em foco não for considerada semanticamente similar à EDU_j pelas heurísticas anteriores, despreza-se essa EDU e busca-se outra, repetindo todo o processo descrito. Determinada uma EDU candidata a compor o sumário, verifica-se se sua inclusão não irá corromper a TC ideal. Essa nova condição pode ser descrita pelos seguintes passos:

- Seja o tamanho do sumário novo denotado por T_1 e o tamanho do sumário pretendido (calculado pela TC) denotado por T_{ideal}. Tem-se em mãos, até o momento, um sumário anterior (tamanho T_0), um sumário novo (tamanho T_1) e o T_{ideal}. A questão, aqui, é decidir qual dos sumários mais se aproxima do desejado.
  - Se T_1 ultrapassar T_{ideal}, comparar com T_0 e escolher, entre o sumário anterior e o novo, aquele cujo tamanho se aproximar mais do ideal. Isso é feito calculando-se a distância entre os sumários, isto é, a diferença de tamanho dos sumários anterior e novo, em relação ao ideal, considerando-se o número de palavras que ultrapassam ou que faltam para atingir o tamanho ideal.
    - Se distância entre T_{ideal} e T_0 for menor que distância entre T_{ideal} e T_1, manter o sumário anterior. Caso contrário, decidir se a construção do

sumário final prossegue com o sumário novo ou se a EDU_j deve ser rebaixada, seguindo o procedimento original do VeinSum, conforme explicitado no capítulo 6.

- Decidindo-se pelo rebaixamento, descarta-se a EDU_j e as EDUs candidatas selecionadas pelo modelo heurístico de correferência, ou seja, volta-se a considerar o sumário anterior, para buscar a próxima EDU mais saliente e retomar os critérios de seleção. Essa nova EDU será a próxima EDU_j a ser incluída no sumário. Portanto, o processo se repete até que um sumário com a melhor aproximação do tamanho ideal seja obtido (podendo ter, apesar de raros os casos, o próprio tamanho ideal).

Tome-se como exemplo um texto-fonte de 100 palavras, com TC de 30%. Nesse caso, o tamanho do sumário ideal (T_{ideal}) deve ser de 30 palavras. Suponha-se que o tamanho do sumário anterior (T_0) seja de 26 palavras e do sumário novo (T_1), de 36 palavras[25]. Neste caso, as respectivas distâncias desses sumários, para o sumário ideal, serão dadas pela diferença entre cada um dos seus tamanhos e o tamanho ideal, como explicado anteriormente. Nesse exemplo, a distância do sumário anterior é de 4 palavras e a do sumário novo é de 6 palavras. Logo, o sumário anterior deve ser mantido. No caso de distância igual entre T_0 e T_1, o sumário anterior deve ser mantido.

- A cada nova EDU incluída, o processo acima deve ser repetido para buscar a clareza referencial também em relação a prováveis anáforas existentes nessa EDU. Ou seja, mesmo que ela não seja a EDU saliente da vez (isto é, mesmo que ela seja proveniente de algum *acc*), o processo deve ser integralmente repetido.
  - Ao esgotar-se a escolha de uma EDU candidata a antecedente da EDU_j dentre as EDUs do *acc*(j), volta-se à lista de EDUs salientes para escolher a próxima a incluir no sumário.

Vale destacar que ao utilizar as heurísticas, não houve a necessidade de uso de nenhuma regra C, pois as regras A e a coincidência de etiquetas foram suficientes para cobrir a totalidade dos casos, apesar de indicarem antecedentes equivocados, algumas vezes. Essas regras, após o processo de filtragem manual, são utilizadas já como heurísticas durante

[25] Taxas de compressão obtidas em cada caso de 26% e 36%, respectivamente.

o processo da construção do novo sumário. Certamente essa forma de defini-las pode acarretar problemas, porém é uma forma válida de se determinar meios de identificação e escolha de EDUs relevantes para um sumário.

O raciocínio geral desta seção é incorporado ao algoritmo descrito a seguir.

6.5 Algoritmo de aprimoramento de seleção de unidades correferentes na SA

Apesar de este ser um trabalho teórico e não visar à implementação do algoritmo abaixo, é importante que ele seja descrito detalhadamente, pois espelha o modelo heurístico teórico descrito na seção anterior.

Passo 1. Para cada EDU_j indicada pela classificação de saliência, recuperar todas as EDUs de *acc*(j), exceto ela própria.
Passo 2. Recuperar CCRs de EDU_j junto ao texto-fonte anotado manualmente com as CCRs de referência. Sejam essas CCRs denotadas por ccrs(j).
Passo 3. Caso nenhum N de nenhum SN da EDU_j esteja em relação de correferência com algum outro SN do texto-fonte, manter EDU_j e respectivo *acc* inteiro no sumário e voltar ao Passo 1. Caso contrário, continuar no Passo 4.
Passo 4. Seja *acc*(j) = {EDU_1, EDU_2,..., EDU_n}. Selecionar EDU_i, tal que EDU_i seja a mais próxima da EDU_j. Como *acc*(j) está sempre em ordem crescente, EDU_i será a EDU que antecede imediatamente EDU_j, em seu *acc*.

1. EDU_i é a que deve ser verificada primeiro, sob a hipótese de que referências anafóricas não estão distantes de seu antecedente.

Passo 5. Recuperar árvore RST do texto-fonte a sumarizar. Se EDU_i for um dos filhos de uma relação de ATTRIBUTION ou de SAME-UNIT, incluir EDU_i no sumário novo. Ir para o Passo 3.
Passo 6. Recuperar as etiquetas semânticas de todos os núcleos de SNs pertencentes à EDU_i.
Passo 7. Recuperar as heurísticas que envolvam o par de etiquetas semânticas de EDU_i e EDU_j, aqui representado pelo par (EDU_i ,EDU_j), seguindo a ordem determinada abaixo:

a. Se alguma regra de associação se aplicar ao par (EDU_i . EDU_j), considerar o par de EDUs semanticamente similares.
b. Caso contrário, buscar a coincidência de etiquetas.

c. Havendo coincidência de etiquetas entre EDU_i e EDU_j, considere o par de EDUs semanticamente similares.
d. Se nem as regras de associação e nem a coincidência de etiquetas se aplicar, utilizar as regras de classificação.
e. Se alguma regra de classificação se aplicar, considere o par de EDUs semanticamente similares.
f. Se EDU_i não for considerada semanticamente similar à EDU_j, pelas heurísticas aplicadas, buscar EDU_{i-1} de *acc*(j) e voltar ao Passo 4.

Passo 8. Verificar se EDU_{i-1} também é considerada semanticamente similar à EDU_j, pela repetição do Passo 7. Incluir todas as EDUs consideradas semanticamente similares à EDU_j do *acc*(j).

Passo 9. Ao incluir uma EDU do *acc*(j), analise também o seu *acc*, pela repetição do processo a partir do Passo 4.

Passo 10. Ao término da análise dos *accs* das EDUs semanticamente similares, calcular o tamanho do sumário novo (T_1).

- Seja o tamanho ideal, indicado por TC desejada, aqui representado por T_{ideal}:
  - Calcular as distâncias do sumário anterior (comprimento T_0) e do sumário novo (comprimento T_1), em número de palavras[26].

dist(sumário anterior)=$|T_{ideal} - T_0|$= 4 e dist(sumário novo)=$|T_{ideal} - T_1|$=6.

a. Se T_1 ultrapassar T_{ideal}, comparar com T_0.
   - Se $|T_{ideal} - T_0| \leq |T_{ideal} - T_1|$, descartar inclusão de EDU_j, sumário anterior = sumário final (tamanho T_0 mantido) e fim do processo.
   - Caso contrário, ficar com sumário novo e fazer sumário novo = sumário anterior (tamanho T_0 recalculado).
   - Se $T_1 = T_{ideal}$, sumário novo = sumário final (tamanho T_1 mantido) e fim do processo.
b. Caso se opte por rebaixar a saliência das EDUs incluídas,

[26] Distâncias calculadas pelo módulo da diferença entre seu tamanho e T_{ideal}.

- Descartar EDU_k do sumário novo, assim como toda EDU_i, $EDU_i \in acc(EDU_k)$ e EDU_i já inclusa no sumário novo.
- Buscar próxima EDU da classificação de saliência (EDU menos saliente que EDU_j), denominando-a de EDU_j.
- Incluir EDU_j no sumário novo.

Passo 11. Se T_1 não ultrapassou T_{ideal} ou houve o rebaixamento da saliência da EDU_j, voltar ao Passo 1, repetindo o mesmo raciocínio.

A reprodução desse algoritmo é ilustrada passo a passo na próxima seção, além das respectivas análises, comparando os novos sumários com os produzidos pelo VeinSum.

6.6 Aplicação do algoritmo, geração e análise dos novos sumários

Dos 50 textos que compõem o *corpus* Summ-it, o modelo heurístico foi aplicado para 11 deles, os quais constam integralmente no Apêndice A. Seus sumários foram gerados manualmente, reproduzindo-se o algoritmo descrito na seção anterior. A construção desses sumários foi manual, pois ao se tratar de um trabalho com base linguística e, portanto, teórico, este mestrado não conta com suporte computacional de qualquer natureza para a implementação do modelo e a obtenção automática dos sumários. Esse suporte, aliás, constituiria um esforço comparável à outra proposta de mestrado, o que evidencia a necessidade de delimitação teórica desta proposta.

Os sumários obtidos foram comparados aos sumários automáticos produzidos pelo VeinSum, também denominados sumários antigos. No que se refere aos problemas dos sumários antigos (explicitados na seção 5.1), propõe-se tratar os que têm a saliência rebaixada e os que corrompem a TC, pois os que não apresentam o antecedente de uma expressão anafórica no texto se devem ao fato de ele não estar presente no *acc* da EDU que contém a anáfora e, portanto, não constituem foco deste trabalho. Já no que se refere às inconsistências dos sumários em relação aos pressupostos do sistema (explicitados na seção 5.2), assumiu-se que o conteúdo do sumário é o correto e os resultados intermediários é que estão incompletos ou equivocados. Somente os sumários que infringiam a regra que determina a inclusão do S de uma relação de ATTRIBUTION, caso o seu N seja determinado pela classificação de saliência, é que foram corrigidos, já que essa também é uma determinação deste modelo heurístico e, se não fossem corrigidos, a comparação dos dados dos dois modelos não seria

justa. Essa correção se deu, simplesmente, pela inclusão do S da relação de ATTRIBUTION ao sumário automático, sem que se excluisse nenhuma outra EDU em detrimento dessa, para que não se corrompesse o resultado original do sistema. Julgou-se que essa correção seria legítima, pois caso o sumário do VeinSum tivesse, ao invés de um S de relação de ATTRIBUTION, outra EDU da classificação de saliência, a comparação que levasse em conta o número de EDUs mais salientes não seria justa, pois é exatamente o foco principal deste trabalho.

Nesta seção, então, são apresentados dois casos em que o modelo de aprimoramento da escolha de EDUs é aplicado levando em consideração o foco inicial e outros casos em que, na tentativa de aplicá-lo, outros fenômenos que poderiam também ser explorados por esse modelo foram descobertos.

6.6.1 Sumários novos que não tiveram a saliência rebaixada

São considerados, nesta seção, os sumários antigos que tiveram a classificação de saliência rebaixada, pois os *accs* das EDUs mais salientes eram longos demais e, com o modelo proposto, seus *accs* puderam ser reduzidos, o que permitiu que EDUs altamente salientes, descartadas no sumário antigo, fossem consideradas nos sumários novos. São dois os casos em que ao aplicar o modelo heurístico foi possível manter a classificação original de saliência no sumário novo. Nas ilustrações que seguem, além dos resultados automáticos do VeinSum, indicados por 'Sum' e 'ResInterm', constam também os resultados da aplicação manual deste modelo heurístico, indicados por 'SumNovo'.

O primeiro caso é o do sumário do texto CIENCIA_2001_19858, ilustrado em Sum13. Observa-se pela comparação do sumário com ResInterm11, que as EDUs 20, 23 e 24 (grifadas em ranking) foram rebaixadas e as EDUs 48 e 2, muito menos salientes, foram incluídas[27].

[27] As EDUs 34 e 36, seguintes às rebaixadas no conjungo ranking, foram incluídas no sumário, pois fazem parte do *acc*(39).

[1] Cientistas do Centro de Estudos de Saclay, na França, parecem ter encontrado o primeiro indício de que a Via Láctea esteja cercada por um verdadeiro campo minado de buracos negros **[2]** - cerca de 1 milhão ao todo, **[4]** A pesquisa, **[7]** é a primeira a constatar que um buraco negro está orbitando o centro da Via Láctea fora do plano do disco galáctico. [11] O objeto em questão [13] já havia sido detectado desde o ano passado, **[14]** mas só agora os cientistas puderam se assegurar de que ele não pertencia ao disco galáctico [34] Um buraco negro nascido na própria Via Láctea que tivesse sido atirado para fora posteriormente, provavelmente pela explosão de uma estrela [36] era justamente o que Mirabel vinha procurando [38] Ainda não foi desta vez que ele achou, **[39]** mas a meta pode não estar longe. **[40]** Seu grupo está atualmente concluindo um segundo estudo, **[48]** A nova pesquisa promete ter impacto ainda maior no meio científico do que o estudo atual.

Sum13. Sumário do texto CIENCIA_2001_19858

<extract method="rank-prune/marcurank/ranksum"
edus="[1,2,4,7,11,13,14,34,36,38,39,40,48]" spine="[4,7,40,14,39,1,34,36,48,2]"
ignored="[24,23,20]"
ranking="[4,7>40>14,39>1,20,23,24,34,36,48>2,8,11,13,19,37,38,44,50>3,9,16,26,27,42,45,49>5,10,12,18,29,35,41,43,46,51>6,15,21,25,28,31,32,47>17,22,30>33]" origlen="539" maxlen="161" len="162" maxrate="32.00" rate="30.06"
delta="0.19"

ResInterm11. Resultados Intermediários do sumário CIENCIA_2001_19858

O sumário já após a aplicação das heurísticas fica como ilustrado em SumNovo1 :

[1] Cientistas <Hprof> do Centro de Estudos de Saclay <org>, na França <civ>, parecem ter encontrado o primeiro indício de que a Via Láctea <Lstar> esteja cercada por um verdadeiro campo minado de buracos negros **[4]** A pesquisa <sem-r>, **[7]** é a primeira a constatar que um buraco negro <Lstar> está orbitando o centro da Via Láctea <Lstar> fora do plano do disco galáctico <Lstar>. [11] O objeto <cc> em questão [13] já havia sido detectado desde o ano passado, **[14]** mas só agora os cientistas <Hprof> puderam se assegurar de que ele não pertencia ao disco galáctico <Lstar>. **[20]** A equipe rastreou a trajetória do buraco negro **[23]** A partir dessas observações eles calcularam a órbita do astro [34] Um buraco negro <Lstar> nascido na própria Via Láctea que tivesse sido atirado para fora posteriormente, provavelmente pela explosão de uma estrela [36] era justamente o que Mirabel vinha procurando. [38] Ainda não foi desta vez **[39]** mas a meta pode não estar longe. **[40]** Seu grupo <HH> está atualmente concluindo um segundo estudo

SumNovo1. Sumário novo do texto CIENCIA_2001_19858

Partindo do conjunto ranking, a primeira saliente é a EDU 4, com única etiqueta semântica **<sem-r>**. Em seu *acc*(4)={1,4}, ou seja, na EDU 1, não há nenhuma etiqueta semântica que possa indicar um possível antecedente para essa possível anáfora, o que não sugere a sua inclusão. Porém a EDU 4 está em relação de SAME-UNIT com a EDU 7, o que já determina a sua inclusão. Essa EDU possui três SNs marcados com **<Lstar>**. Analisando o *acc*(7)={1,4,7}, como a EDU 4 já foi incluída e não possui etiquetas que possam levar aos possíveis antecedentes das possíveis anáforas em 7, analisa-se a EDU 1. A inclusão dessa EDU se faz necessária agora, pois, apesar de nenhuma regra A se aplicar, há coincidência de etiquetas entre as duas EDUs. Porém, das três possíveis anáforas da EDU 7, somente uma é realmente uma menção anafórica e as outras duas se tratam de primeiras menções de CCRs. Essa anáfora, no entanto, tem o seu antecedente corretamente indicado pela inclusão da EDU 1, apesar de ser uma simples repetição lexical.

Calcula-se a TC e como ainda não se atingiu a ideal, o processo se repete. A próxima saliente é a EDU 7, já incluída no sumário. Analisa-se, então, a EDU 40, com única possível anáfora marcada com **<HH>**. Analisando o *acc*(40)={4,7,39,40} verifica-se que a EDU 39 não possui nenhuma etiqueta semântica e deve, então, ser descartada. Como as EDUs 7 e 4 já constam no sumário e a única anáfora da EDU 40, marcada com **<HH>**, tem o seu antecedente já explícito na EDU 1, pela correta aplicação da regra A2 (**<HH>, <Hprof>**), calcula-se a TC e o processo se repete, já que ainda não se atingiu os 30%.

Seguindo a próxima saliente, a EDU 14, encontram-se duas possíveis anáforas **<Hprof>** e **<Lstar>**. Pela análise do *acc*(14)={4,7,11,13,14}, a EDU 13, mais próxima, deve ser desconsiderada já que não possui etiqueta semântica alguma. A EDU 11, no entanto, possui um candidato a antecedente para **<Lstar>**, apontado pela regra A23 (**<Lstar>**,**<cc>**) e essa EDU deve, portanto, ser incluída, apesar de não indicar o antecedente correto da anáfora marcada com **<Lstar>**. Com a inclusão da EDU 11, a EDU 13 deve também ser incluída, por estar ligada à EDU 11 por uma relação de SAME-UNIT, mesmo tendo sido descartada no passo anterior. O *acc*(11)={4,7,11} não determina a inclusão de nenhuma outra EDU além das que já constam no sumário e nem o *acc*(13)={4,7,11,13}. A outra possível anáfora constante em 14 tem o seu antecedente já no texto, proveniente da coincidência de etiquetas com a EDU 1, do *acc*(7). Novamente a TC é calculada e a próxima de ranking é incluída.

A EDU 39, não possui etiqueta semântica alguma, porém, não pode ser desprezada pois é a próxima saliente. Isso justifica a sua inclusão, porém não há como descartar EDUs do *acc*(39)={4,7,34,36,38,39}, já que não há possível anáfora nessa EDU e, portanto, não faz sentido buscar possíveis antecedentes. Calcula-se a TC novamente e como o sumário ainda não atingiu o tamanho ideal, o processo se repete.

A EDU 1, próxima saliente, já faz parte do sumário e a EDU 20 é, então, analisada. Nessa EDU encontram-se duas possíveis anáforas, marcadas com **<HH>** e **<Lstar>**. Em seu *acc*(20)={4,7,14,19,20}, a EDU 19 não apresenta etiqueta semântica, o que indica a sua exclusão. As demais EDUs desse *acc* já estão incluídas no sumário e as anáforas da EDU 20 já têm seus antecedentes explícitos. A anáfora marcada com **<HH>** encontra o seu antecedente mais próximo na EDU 14, marcado com **<Hprof>** e identificado pela regra A2 (**<HH>**,**<Hprof>**). Ainda, o antecedente mais completo dessa menção anafórica encontra-se também no texto, marcado com etiqueta **<Hprof>** e pertencente à EDU 1 e a CCR é, portanto, preservada. A anáfora **<Lstar>** também tem o seu antecedente explícito na EDU 7, através da coincidência de etiquetas.

Calcula-se novamente a TC e inclui-se a próxima saliente, a EDU 23, que possui duas possíveis anáforas, marcadas com **<act-d>** e **<Lstar>**. Pela análise do seu *acc*(23)={4,7,14,19,20,23}, verifica-se que a possível anáfora **<Lstar>** tem um antecedente indicado corretamente através da coincidência de etiqueta com a EDU 20. Já a EDU 19 deve ser excluída por não conter nenhuma etiqueta similar ou igual às das possíveis anáforas em

23. A possível anáfora marcada com **<act-d>** em 23, no entanto, não tem o seu antecedente explícito, pois ele se encontra na EDU 21 e essa não faz parte do sumário.

O novo cálculo de TC revela que o sumário já atinge a TC ideal e o processo é, então, finalizado.

Ao analisar o sumário novo, verifica-se uma quebra de CCR na EDU 36, na qual menciona-se Mirabel e, pelo contexto do sumário, não é possível recuperar o seu antecedente. Essa quebra está presente também no sumário antigo e se deve ao fato de o antecedente 'Félix Mirabel', constante da EDU 30, não estar presente no *acc*(36)={4,7,34,36} e nem recursivamente no *acc*(34)={4,7,34}. Esse é um exemplo de problema que pode ser proveniente tanto de erros no algoritmo da VT quanto de árvores RST mal estruturadas.

Observa-se também que, das 3 EDUs rebaixadas no sumário antigo, 2 delas são contempladas no sumário novo e, portanto, esse foi construído com EDUs mais salientes que as consideradas no sumário antigo. Além disso, no sumário novo foi considerada a TC ideal de 30%, 2% a menos que a do sumário antigo, o que caracteriza mais uma vantagem do modelo proposto.

Outro caso em que houve rebaixamento da saliência no sumário antigo é o do texto CIENCIA_2000_17101, constante de Sum14. Observa-se, pela comparação de ranking com ignored, em ResInterm12, que a classificação de saliência foi rebaixada: as EDUs grifadas nos dois conjuntos seriam candidatas a inclusão no sumário, pois são consideradas altamente salientes pelo conjunto ranking. Foram, porém, descartadas, pois tinham os *accs* longos demais, a ponto de corromper a TC e, por isso, se encontram no conjunto ignored. Consequentemente, Sum14 foi construído com EDUs muito menos salientes, como se pode observar pela posição que a EDU 5 ocupa em ranking, pois essa somente foi incluída no sumário após o descarte de quinze EDUs mais relevantes que ela.

[1] 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, **[2]** que a comunidade científica internacional pode alterar a Declaração de Helsinque, sobre ética em pesquisas com humanos. **[3]** Em estudos no Terceiro Mundo, os cientistas se desobrigariam de fornecer aos doentes o melhor tratamento médico conhecido. **[4]** A proposta, **[5]** a ser discutida, **[6]** dá aos pesquisados o direito de receber terapia dada pelo governo de seu país **[7]** - que pode ser nenhuma.

Sum14. Sumário do texto CIENCIA_2000_17101

<extract method="rank-prune/marcurank/ranksum" edus="[1,2,3,4,5,6,7]"
spine="[3,2,1,4,6,7,5]"
ignored="[31,33,30,27,25,23,32,29,24,11,8,28,26,20,19,15,10,22,14,13,12,9,21,18,17,16]"
ranking="[3>16,17,18,21>2,9,12>1,4,6,13>7,14,22>10,15,19,20,26,28>5,8,11,24,29,32>23,25>27,30,33>31]"
origlen="343" maxlen="102" len="87" maxrate="30.00" rate="25.36" delta="-18.28"

ResInterm12. Resultados Intermediários do sumário CIENCIA_2000_17101

Em SumNovo2, no entanto, não há nenhuma EDU saliente descartada e o sumário as contempla em maior número, portanto.

[1] O presidente <Hprof> da Comissão Nacional de Ética em Pesquisa <org>, William Saad Hossne <hum>, disse ontem, na 52ª Reunião Anual da Sociedade Brasileira para o Progresso da Ciência, [2] que a comunidade científica internacional pode alterar a Declaração de Helsinque <sem-r>, sobre ética <domain> em pesquisas com humanos. **[3]** Em estudos no Terceiro Mundo <civ>, os cientistas <Hprof> se desobrigariam de fornecer aos doentes <Hsick> o melhor tratamento <activity> médico conhecido.
[9] em que cientistas <Hprof> deram a grávidas <Hattr> com HIV <meta> um regime <activity> de AZT <cm-rem> mais breve do que o recomendado. [15] Os pesquisadores <Hprof> argumentaram **[16]** que as mulheres <Hattr> que não receberam AZT <cm-rem> não o teriam recebido, de qualquer forma.

SumNovo2. Sumário novo do texto CIENCIA_2000_17101

O primeiro passo para a sua construção foi a inclusão da EDU 3, primeira de ranking, e suas possíveis anáforas são consideradas tendo em vista os núcleos de todos os SNs que compõem essa EDU. Esses recebem etiquetas semânticas somente se estiverem em relação de correferência com algum outro SN do texto. As possíveis anáforas, nesse caso, recebem etiquetas **<civ>**, **<Hprof>**, **<Hsick>** e **<activity>**, respectivamente.

A seguir, foi analisado o *acc*(3)={2,3}, a fim de encontrar alguma etiqueta que remeta a algum possível antecedente. A regra A7 (**<sem-r>**,**<activity>**) pode ser aplicada neste caso, pois as EDUs 2 e 3 contêm núcleos de SNs com essas etiquetas e, portanto, a EDU 2 deve ser incluída no sumário, o que indica que, para a possível anáfora **<activity>**, um potencial antecedente é **<sem-r>**. Essa etiqueta, no entanto, introduz um SN não correferente com o item marcado com **<activity>**, que se trata do antecedente mais completo de outra CCR e, portanto, não causa nenhuma quebra no sumário. No entanto, nenhuma anáfora de 3 tem seu antecedente explicitado com a inclusão da EDU 2. Porém, ao incluir a EDU 2, é preciso analisar também o *acc*(2)={1,2}. Ainda levando em conta as possíveis anáforas da EDU 3 e analisando a EDU 1, pode-se aplicar a regra A1 (**<hum>**,**<Hprof>**), que indica que **<hum>** é um antecedente possível caso **<Hprof>** seja uma expressão anafórica, e isso pressupõe a inclusão da EDU 1. O uso dessa regra também não é adequado, pois o antecedente indicado por ela não é o correto, ou seja, o antecedente indicado por A1 'William Saad Hossne' não constitui o real antecedente da anáfora 'cientistas', na EDU 3. Ainda que a coincidência de etiquetas **<Hprof>** das EDU 3 e 1 fosse considerada, também não indicaria o antecedente correto, pois 'o

presidente' jamais seria correferente com 'os cientistas'. Nesses casos específicos, nem informações de gênero e número podem contribuir, já que tanto a regra, quanto a coincidência de etiquetas já levam a antecedentes equivocados. Porém, comparando-se o sumário do VeinSum com esse, não há diferença de conteúdo até o presente momento.

Escolhidas as EDUs do *acc*(3) que devem ser incluídas, verifica-se o tamanho do sumário e como ainda não se atingiu a TC ideal, escolhe-se a próxima EDU saliente segundo o ranking, que é a 16, e o processo se repete.

No caso dessa EDU, as possíveis anáforas são **<Hattr>** e **<cm-rem>**. Busca-se, então, no *acc*(16)={3,9,12,15,16} a sua mais próxima e, como a EDU 15 está ligada à EDU 16 por uma relação de ATTRIBUTION, ela deve ser incluída incondicionalmente. Ao incluir a EDU 15, busca-se etiquetas similares ou iguais no *acc*(15)={3,9,12,15}, pois esse pode, por sua vez, conter uma possível anáfora. Recai-se sobre a EDU 12, a mais próxima de 15. Essa EDU, no entanto, contém um possível antecedente somente para uma das possíveis anáforas de 16, indicado pela coincidência de etiquetas. Se comparada à EDU 9, na qual encontram-se possíveis antecedentes para as duas possíveis anáforas de 16 e ainda a introduzida em 15, todas por coincidência de etiquetas, optou-se por descartar a EDU 12 e manter somente a EDU 9, por essa indicar mais possibilidades de antecedentes. Ao descartar a EDU 12, não se torna necessária a análise de seu *acc*. Ainda no *acc*(15), analisa-se a EDU 9 e o *acc*(9)={3,8,9}. Observa-se que na EDU 9 existem possíveis antecedentes para as anáforas possíveis de 16. Apesar de nenhuma regra A se aplicar, encontram-se etiquetas coincidentes nessas duas EDUs, o que requer a inclusão de 9. Em seu *acc*, no entanto, a EDU 8 não apresenta nenhuma etiqueta semântica correferente com as possíveis anáforas de 9 ou de 16, o que sugere que essa EDU não seja incluída. Como a EDU 3, pertencente ao *acc*(9), já consta no sumário, calcula-se a TC, então, a fim de verificar se a próxima EDU saliente deve ou não ser incluída. Neste caso, o sumário já atinge a TC ideal e o processo é finalizado.

A análise linguística da coincidência de etiquetas revela que o item lexical 'AZT', marcado com **<cm-rem>** teve o seu antecedente indicado corretamentetamente pela etiqueta **<cm-rem>**. Um ponto importante a ser observado é que anáfora e antecedente são casos de repetição lexical. Nesses casos, o tratamento deveria ser diferente de quando se tem a mesma etiqueta semântica para itens lexicais diferentes, pois na repetição lexical o significado está autocontido. Por exemplo, na ocorrência de uma CCR em que o antecedente é 'Júpiter' e uma anáfora é o 'planeta', os dois itens lexicais recebem a etiqueta **<Lstar>**. Porém, esse caso é problemático se o antecedente não se encontra no sumário, pois a anáfora 'o planeta' pode se

referir a qualquer planeta existente. Já no caso da repetição lexical, mesmo que o antecedente não esteja presente no texto, sua presença é dispensável para o entendimento da mensagem.

Outro caso de coincidência de etiquetas identificado corretamente é o da etiqueta **<Hprof>**. Há uma preservação parcial da CCR, porém a anáfora 'pesquisadores', na EDU 15, tem seu antecedente mais completo marcado como 'cientistas' no *corpus* de referência (anotado manualmente) o que não caracteriza o antecedente mais completo e sim outro termo anafórico da mesma CCR. Seu antecedente mais completo deveria ser 'comunidade científica internacional' e isso caracteriza, portanto, erro na anotação manual, o que não possibilitou a aplicação da regra A2 **(<HH>**, **<Hprof>**). Se essa tivesse sido aplicada, o antecedente seria indicado corretamente, mas como 'comunidade científica internacional' não é marcado como correferente a nenhum outro SN do texto e, portanto, não apresenta etiqueta semântica, essa relação não foi identificada. Isso não comprova a premissa (i), pois o *corpus* anotado com as informações de correferência também apresenta inconsistências e deve ser revisto.

A anáfora que não teve seu antecedente apontado corretamente foi etiquetada com **<Hattr>**, que tinha dois possíveis antecedentes, em EDUs diferentes, com etiquetas coincidentes. A escolha da EDU que continha o maior número de candidatos a antecedente prejudicou a identificação correta do mesmo, que estava na EDU 12, descartada. Casos como esse, podem ser problemáticos para esta proposta, pois a diferença entre 'grávidas com HIV que receberam um regime de AZT mais breve que o recomendado' e 'mulheres (grávidas) que não receberam AZT' não está em N e as regras geradas neste trabalho somente levam N em consideração. Nesse caso, 'grávidas', N do primeiro SN e 'mulheres', N do segundo SN, recebem a mesma etiqueta e, assim, por esta proposta, esses SNs passariam a ser correferentes, o que não é verdade. Essa constatação, portanto, não comprova a hipótese (iv), pois a aplicação das heurísticas pode comprometer o encadeamento referencial estabelecido no texto-fonte. Entretanto, isso ocorreu em apenas um caso dos 11 textos analisados e pelo contexto do sumário é possível identificar a diferença entre os SNs, o que não caracteriza um problema de coerência.

Um problema deste sumário caracteriza-se pela exclusão do S da relação retórica de CIRCUMSTANCE, pois o seu N preservado causa uma sensação de mensagem incompleta no leitor, como ilustrado na Figura 20. Porém, o conteúdo do S somente traz um detalhe, que para fins de SA não seria problema descartá-lo. O problema maior aqui é o trecho 'em que',

preservado no sumário sem que se indique a circunstância. Se esse trecho não constasse no sumário, essa sensação de incompletude não ocorreria.

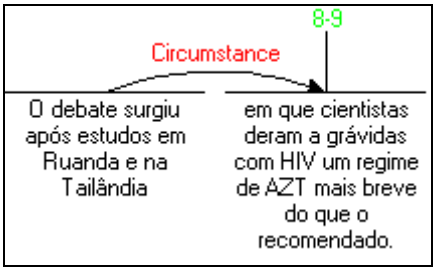


Figura 20. Ilustração da mensagem completa dada pela relação de CIRCUMSTANCE

Apesar de algumas das heurísticas ou etiquetas coincidentes levarem a antecedentes equivocados em alguns casos, o sumário não apresenta nenhum problema de clareza referencial. Além disso, não teve nenhuma EDU saliente descartada e sua TC, que no sumário automático era de 27%, passa a ser 30%, a considerada ideal.

A proposta foi então aplicada aos demais casos e resultados diferentes foram obtidos, como os descritos a seguir.

6.6.2 Sumários novos que apresentam TC mais próxima da ideal

Dentre os casos que não apresentaram rebaixamento da saliência, observaram-se algumas circunstâncias em que esses podiam ter sua TC reduzida e, portanto, mais próxima da ideal, tais como: i) redução do tamanho do sumário quando sua TC ultrapassa a ideal através do descarte de EDUs provenintes de *accs* e; ii) sumários construídos apenas com EDUs altamente salientes ou S de relação de ATTRIBUTION que tiveram sua TC reduzida no sumário novo.

Redução de TC do sumário novo pelo descarte de EDUs provenientes de *accs*

Um exemplo desse caso é o do sumário do texto CIENCIA_2000_17088, ilustrado em Sum15. Observa-se no item rate, em ResInterm13, que o tamanho do sumário ultrapassou em 9,22% a TC desejada. Com a utilização do modelo heurístico, é possível reduzir o *acc* de uma EDU saliente e chegar a uma TC de 34%, como ilustra SumNovo3.

[1] Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de uma nova espécie de dinossauro no Brasil. [4] Batizado de Santanaraptor placidus, **[5]** o fóssil é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele. **[11]** Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o Santanaraptor ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex, [19] O exemplar de Santanaraptor encontrado pela equipe carioca foi desenterrado em 1991, [20] mas a montagem do fóssil só foi concluída nove anos mais tarde. [21] Tudo o que sobrou dele foram as patas e partes da cauda e da bacia, **[22]** mas os pesquisadores conseguiram estimar que o bicho fosse um filhote de 1,5 metro de altura.

Sum15. Sumário do texto CIENCIA_2000_17088

```
<extract method="rank-prune/marcurank/ranksum" edus="[1,4,5,11,19,20,21,22]"
spine="[1,5]" ignored="[]"
ranking="[1>5,11,22>2>3,10,16,23>6,12,17,20,26>4,7,13,18,19,21,25,27>8,14,24>
9,15]" origlen="357" maxlen="107" len="140" maxrate="30.00" rate="39.22"
delta="23.50">
```

ResInterm13. Resultados Intermediários do sumário CIENCIA_2000_17088

O sumário novo teve somente uma EDU proveniente de *acc* descartada, comparado ao sumário antigo – a EDU 21, porém essa já foi responsável pela redução de 5% no tamanho do sumário. Observe o processo de construção do sumário novo a seguir.

[1] Pesquisadores <Hprof> do Museu Nacional <org> do Rio de Janeiro <civ> anunciaram ontem a descoberta de uma nova espécie <meta> de dinossauro <Azo> no Brasil <civ>. [4] Batizado de Santanaraptor placidus <meta>, **[5]** o fóssil <cc-stone> é o único a ser encontrado no país <Lciv> com restos de tecido <an> mole, como fibras musculares, vasos sanguíneos e pele. **[11]** Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o Santanaraptor <meta> ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex <meta>, [19] O exemplar de Santanaraptor <meta> encontrado pela equipe <HH> carioca foi desenterrado em 1991, [20] mas a montagem do fóssil <cc-stone> só foi concluída nove anos mais tarde. **[22]** mas os pesquisadores <Hprof> conseguiram estimar que o bicho <Azo> fosse um filhote de 1,5 metro de altura.

SumNovo3. Sumário novo do texto CIENCIA_2000_17088

A primeira EDU marcada por ranking é a EDU 1, o que pressupõe a sua inclusão no sumário sem a necessidade de análise do *acc*(1), pois no caso da primeira EDU do texto, o seu *acc* somente compreende ela mesmo, já que não há nenhum segmento textual antes dele.

Analisa-se, então a próxima do ranking, a EDU 5. Dentre as suas possíveis anáforas, a indicada pela etiqueta **<cc-stone>** tem o seu antecedente indicado corretamente pela regra A98 (**<cc-stone>**,**<meta>**), proveniente da EDU 4 do *acc*(5)={1,4,5}. Isso requer a sua inclusão e o *acc*(4)={1,4} somente compreende EDUs já incluídas no sumário. Ainda nas anáforas de 5, **<Lciv>** tem seu antecedente **<civ>** indicado corretamente pela regra A3 (**<Lciv>**,**<civ>**), já na EDU1. Já para a etiqueta **<an>**, também pertencente à EDU 5, a regra A30 indica que essa etiqueta coocorre com **<meta>** e já existem duas EDUs com essa etiqueta no sumário. Porém o item lexical marcado com **<an>** não configura uma anáfora, mas sim o antecedente mais completo de outra CCR do texto. Assim, como a inclusão dessas EDUs não foi determinada por essa etiqueta, não há diferença no sumário.

A próxima EDU saliente segundo o ranking é a EDU 11. Suas possíveis anáforas são indicadas por **<meta>**. Analisando o *acc*(11)={1,5,11} verifica-se que a regra A30 (**<an>**,**<meta>**) pode ser aplicada, o que sugere, incorretamente, que o antecedente de uma das anáforas marcadas com **<meta>** é 'tecido mole', sendo que esse é parte de outra CCR. Essa EDU, no entanto, foi determinada pelo Modelo de Saliência e já consta no sumário, o que não compromete a clareza referencial já que o antecedente correto de **<meta>** já consta na EDU1 – 'nova espécie de dinossauro', com informação complementar na EDU4 – 'Santanaraptor

placidus'. O outro item lexical marcado com **<meta>** nessa mesma EDU não configura uma anáfora e, portanto, não introduz quebra de CCR.

Em seguida, a análise recai sobre a EDU 22, próxima do ranking, e seu *acc*(22)={1,5,11,20,21,22}. Suas possíveis anáforas são marcadas com **<Hprof>** e **<Azo>**. Na EDU 21 não há etiqueta semântica alguma, o que determina a sua exclusão. Já na EDU 20 encontra-se um possível antecedente para **<Azo>**, etiquetado como **<cc-stone>**, segundo A97, o qual é determinado corretamente. Ao incluir a EDU 20, analisa-se também o seu *acc*(20)={1,5,11,19,20}. Ao analisar a EDU 19, é possível encontrar o antecedente correto para a anáfora **<Hprof>**, que segundo A2, é **<HH>**. As demais EDUs do *acc*(19)={1,5,11,19} já fazem parte do sumário e volta-se para a análise do *acc*(22). As demais EDUs desse *acc* também já estão incluídas no sumário e não há como chegar na TC ideal a não ser que se rebaixe a EDU 22. Caso o VeinSum tivesse rebaixado a EDU 22, as EDUs 2 e 3 teriam sido incluídas no sumário e esse ficaria com exatos 30%. Porém, como o sistema não rebaixou a EDU, julgou-se que não seria conveniente rebaixá-la também. Assim, com a redução do *acc*(22), a TC do sumário novo chega aos 34%, sendo 5% menor que a apresentada pelo sumário automático.

A EDU excluída era um S da relação retórica de CONCESSION, como ilustra a Figura 21. Seu conteúdo não é significativo em termos de sumarização, pois constitui detalhe somente. No entanto, a mensagem veiculada pelo N causa no leitor uma sensação de mensagem incompleta devido, exclusivamente, à presença da conjunção 'mas', que pressupõe algo que possa complementar a ideia de concessão.

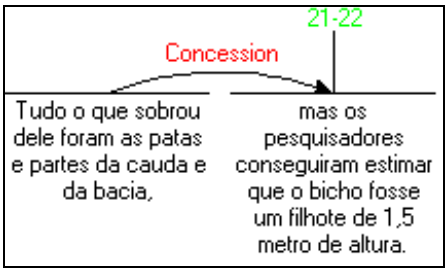


Figura 21. Ilustração da mensagem completa dada pela relação CONCESSION

Apesar disso, não há nenhuma quebra de clareza referencial e nenhuma EDU foi rebaixada, assim como no sumário antigo também não havia. A contribuição aqui foi a redução da TC, que chegou mais próxima da ideal.

Outro exemplo da possibilidade de redução de TC é o sumário do texto CIENCIA_2000_17108, ilustrado em Sum16, com seus respectivos resultados intermediários constantes de ResInterm14.

[1] Um ser que invade corpos e domina a mente alheia, [3] não é mero personagem de ficção. [4] Para uma aranha da Costa Rica, essa criatura existe. [6] Apesar do nome [8] o tal invasor de corpos é só uma vespa. [9] O biólogo William Eberhard, da Universidade da Costa Rica, descobriu **[10]** que as larvas desse inseto **[12]** provocam mudanças no comportamento da hospedeira. **[36]** "É uma descoberta e tanto", [37] disse o psicólogo César Ades, da USP, especialista em comportamento de aranhas **[38]** "É a primeira vez que se vê uma interação química tão complexa entre parasita e hospedeiro", [39] afirmou. **[40]** A exploração alheia não tem limites.

Sum16. Sumário do texto CIENCIA_2000_17108

Esse sumário foi corrigido pela inclusão das EDUs 37 e 39, já que essas estão ligadas à EDU saliente incluída por constituírem S de uma relação de ATTRIBUTION. A TC do sumário sem correção era de 35% e passou a ser de 39%. Não se excluiu nenhuma EDU a fim de não corromper o sumário original e, além disso, porque não é claro o processo que o VeinSum segue com relação à finalização do sumário, já que em muitos casos a TC é corrompida sem necessidade.

```
<extract method="rank-prune/marcurank/ranksum"
edus="[1,3,4,6,8,9,10,12,36,38,40,41]" spine="[10,12,36,38,4,6,8,40,1,3,41,9]"
ignored="[27,21,11,7,35,32,31,28,25,22,19,15,14,5,2,39,37,34,33,30,29,26,24,23,2
0,18,17,16,13]"
ranking="[10,12,36,38,4,6,8,40,1,3,13,16,17,18,20,23,24,26,29,30,33,34,41,9,37,3
9,2,5,14,15,19,22,25,28,31,32,35,7,11,21,27]" origlen="340" maxlen="102"
len="102" maxrate="30.00" rate="30.00" delta="0.00">
```

ResInterm14. Resultados Intermediários do sumário CIENCIA_2000_17108

O sumário novo, ilustrado em SumNovo4, tem a TC reduzida.

[1] Um ser <aent> que invade corpos <anmov> e domina a mente <anorg> alheia, [3] não é mero personagem de ficção. [4] Para uma aranha da Costa Rica, essa criatura <Azo> existe. [6] Apesar do nome [8] o tal invasor <aent> de corpos <anmov> é só uma vespa <aent>. [9] O biólogo <Hprof> William Eberhard <hum>, da Universidade da Costa Rica, descobriu **[10]** que as larvas desse inseto <aent>, **[12]** provocam mudanças no comportamento da hospedeira <Azo>. **[36]** "É uma descoberta e tanto", [37] disse o psicólogo César Ades <hum>, da USP, especialista em comportamento de aranhas.

SumNovo4. Sumário novo do texto CIENCIA_2000_17108

A primeira EDU de ranking a ser introduzida é a EDU 10, com uma possível anáfora etiquetada com **<aent>**. A EDU 9 está ligada à EDU 10 através de uma relação de ATTRIBUTION e deve, portanto, ser incluída no sumário. Seu *acc*(9)={4,6,8,9} deve, então, ser analisado. As possíveis anáforas da EDU 9 estão marcadas com **<Hprof>** e **<hum>** e como nenhuma EDU do seu *acc* é considerada semanticamente similar a ela, o conjunto todo deve ser incluído no sumário. Entretanto, as expressões marcadas semanticamente na EDU 9 não constituem anáforas, mas sim o antecedente mais completo de outra CCR do texto. A inclusão de todas as EDUs do *acc*(9), pressupõe a análise do *acc* de cada uma delas. Inicia-se pela análise da EDU 8. Dentre as suas possíveis anáforas, duas delas estão marcadas com **<aent>** e uma com **<anmov>**. O *acc*(8)={1,3,4,6,8} determina a inclusão da EDU 1, já que nenhuma regra ou coincidência de etiquetas se aplica às EDUs já constantes do sumário e a EDU 1 indica o antecedente correto para a anáfora. Os demais *accs*, *acc*(4)={1,3,4}, *acc*(3)={1,3} e o *acc*(1)={1} não necessitam de análise já que todas as EDUs já constam no sumário. Volta-se para a análise do *acc*(10)={4,6,8,9,10} e verifica-se também que todas as suas EDUs já constam no sumário. Calcula-se a TC e a análise recai sobre a próxima EDU saliente, a EDU 12. Essa contém uma possível anáfora marcada com **<Azo>**. Seu *acc*(12)={4,6,8,9,10,12}, no entanto, não apresenta nenhuma EDU a ser analisada, já que todas já constam do sumário. Calcula-se a TC novamente.

A próxima saliente é a EDU 36, com respectivo *acc*(36)={10,12,36}. Essa EDU não possui nenhuma etiqueta semântica, mas deve ser considerada no sumário, pois se trata da próxima EDU saliente na classificação. A EDU 37 deve ser também incluída, já que constitui S da relação de ATTRIBUTION e está ligada à EDU 36 por esse motivo. O *acc*(36) e o *acc*(37)={10,12,36,37} somente indicam EDUs que já constam no sumário e, portanto,

nenhuma outra EDU deve ser incluída até o momento. Calcula-se a TC e essa já atinge os 31%. O processo é, então, finalizado.

Esse sumário se aplica a esse caso, pois o sumário antigo foi o considerado e já apresentava TC de 35%.

Essa redução de tamanho também ocorre para o texto CIENCIA_2000_17109, ilustrado em Sum17. Sua TC é de 40,57%, como se observa no item rate em ResInterm15.

[1] Foi dado o primeiro passo para a diminuição das filas de espera para transplante de fígado. **[2]** Cientistas britânicos detectaram, em adultos, a produção de células hepáticas a partir de células-tronco da medula óssea. [16] Pesquisas em camundongos haviam mostrado que células-tronco da medula óssea poderiam originar células hepáticas, além das sanguíneas. [17] Para descobrir se o mesmo acontecia em seres humanos, **[18]** os pesquisadores analisaram células do fígado de mulheres que haviam sofrido um transplante de medula óssea, **[20]** A análise do DNA dessas células mostrou que elas continham o cromossomo Y **[23]** O resultado deverá contribuir para o desenvolvimento de tecidos humanos para uso terapêutico

Sum17. Sumário do texto CIENCIA_2000_17109

```
<extract method="rank-prune/marcurank/ranksum" edus="[1,2,16,17,18,20,23]"
spine="[2,23,18]" ignored="[]"
ranking="[2>23>18,20,24>10,16>5,7,9,12,14>1,3,11,15,17,22>4,19,21>6,13>8]"
origlen="281" maxlen="84" len="114" maxrate="30.00" rate="40.57"
delta="26.05">
```

ResInterm15. Resultados Intermediários do sumário CIENCIA_2000_17109

Com o novo sumário, ilustrado em SumNovo5, essa taxa cai para 32%.

[1] Foi dado o primeiro passo <act-d> para a diminuição <event> das filas <coll> de espera <activity> para transplante <act> de fígado <anorg>. **[2]** Cientistas <Hprof> britânicos <Hnat> detectaram, em adultos, a produção <activity> de células hepáticas <an> a partir de células-tronco <an> da medula óssea <anorg>. [16] Pesquisas em camundongos haviam mostrado que células-tronco <an> da medula óssea <anorg> poderiam originar células hepáticas <an>, além das sanguíneas. **[18]** os pesquisadores <Hprof> analisaram células <an> do fígado <anorg> de mulheres <Hattr> que haviam sofrido um transplante <act> de medula óssea <anorg>, **[23]** O resultado deverá contribuir para o desenvolvimento de tecidos humanos para uso terapêutico

SumNovo5. Sumário novo do texto CIENCIA_2000_17109

Dentre as possíveis anáforas da EDU 2, primeira do ranking, as que podem ter seus antecedentes explícitos pela análise do *acc*(2)={1,2} são: **<activity>** e **<anorg>**. A etiqueta **<activity>** pode ter seu antecedente identificado pela regra A28 (**<act>**,**<activity>**) ou pela regra A118 (**<event>**,**<activity>**), além da coincidência de etiquetas nas duas EDUs, o que obriga a inclusão da EDU1. Nesse caso, **<activity>** tem o seu antecedente identificado corretamente pela regra A118. No entanto, a regra somente leva o N em consideração e a correspondência aqui se dá através do SN todo, ou seja, a 'produção de células hepáticas a partir de células-tronco da medula óssea' tem como antecedente marcado no *corpus* de referência 'o primeiro passo para a diminuição das filas de espera para transplante de fígado'. Já para a possível anáfora marcada com **<anorg>**, ainda na EDU 2, encontra-se um candidato a antecedente através da coincidência de etiquetas. No entanto, o SN marcado com **<anorg>** na EDU 2 não se caracteriza como uma anáfora, mas sim como o antecedente mais completo de outra CCR do texto.

Calcula-se a TC e inclui-se a próxima saliente, já que ainda não se atingiu os 30%. A EDU 23, no entanto, não tem nenhum de seus SNs correferente com outro do texto e, portanto, nenhuma etiqueta semântica presente. Essa EDU, no entanto, não pode ser excluída, pois é considerada altamente saliente e a análise de seu *acc*(23)={2,23} não prevê a inclusão de nenhuma outra EDU, já que a 2 também já consta no sumário. Calcula-se novamente a TC.

A EDU 18, por sua vez, contém seis possíveis anáforas. Analisando o *acc*(18)={2,16,17,18} percebe-se que a EDU 17 não deve ser considerada já que não possui nenhuma etiqueta semântica e, consequentemente, seu *acc* não é considerado. A EDU 16, no

entanto, contém possíveis antecedentes para três das seis anáforas da EDU 18. A etiqueta **<an>** não tem seu antecedente apontado por nenhuma regra A, porém a EDU 16 apresenta dois SNs marcados com a mesma etiqueta dessa anáfora e um deles indica o antecedente correto, preservando a cadeia. Duas outras possíveis anáforas de 18 são marcadas com **<anorg>**, que não consta em nenhuma regra A ou C, mas que apresenta uma etiqueta coincidente na EDU 16, a qual indica o antecedente correto. O *acc*(16)={2,16} já foi incluído no sumário e, portanto, volta-se para a análise da EDU 18. Sua anáfora etiquetada com **<Hprof>**, tem seu antecedente apontado corretamente pela coincidência de etiquetas na EDU 2, parte do *acc*(18), e já incluída no sumário. Calcula-se a TC e essa está com 32%, ou seja, 8% menor do que a considerada no sumário automático.

O processo de construção do sumário novo até esse momento é exatamente igual ao do sumário antigo, porém como a TC já é atingida não se considera a EDU 20, próxima saliente. No sumário antigo, no entanto, ela é incluída, o que faz com que a TC aumente de 32% para 40% e, além disso, sua inclusão introduz uma mensagem deturpada, já que a ela indica que 'as células do fígado das mulheres transplantadas continham o cromossomo Y'. Porém, pelo texto-fonte, constante do Apêndice A, é possível recuperar a mensagem integral e isso somente ocorre em casos cujos doadores eram homens e não, genericamente, como consta no sumário antigo.

O sumário novo, no entanto, não apresenta essa deturpação da mensagem e também não corrompe tanto a TC como o antigo. Além disso, não apresenta nenhuma quebra de clareza referencial. Esses são indícios de que esse sumário é melhor estruturado que o antigo.

Sumários construídos apenas com EDUs salientes e Ss de relação de ATTRIBUTION que tiveram a TC reduzida

O sumário do texto CIENCIA_2000_6389, ilustrado em Sum18, é um caso no qual nenhuma EDU pode ser dispensada, pois ou são indicadas como altamente salientes ou são S da relação de ATTRIBUTION, como se pode observar em ResInterm16.

[1] A discussão sobre a biotecnologia nacional está enviesada, **[5]** Guerra citou a micropropagação de vegetais **[10]** como exemplo de biotecnologia de baixo custo. [12] Para o agrônomo, **[13]** o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local, como o cultivo de plantas com a capacidade de captar certos elementos presentes na terra. [14] O presidente da Embrapa [16] Alberto Portugal, salientou **[17]** que a empresa busca soluções para os problemas da agricultura nacional. [19] Portugal disse **[20]** que os agronegócios correspondem a 25% do PIB brasileiro e **[21]** que a biotecnologia é fundamental.

Sum18. Sumário do texto CIENCIA_2000_6389

```
<extract method="rank-prune/marcurank/ranksum"
edus="[1,5,10,12,13,14,16,17,19,20,21]" spine="[1,5]" ignored="[]"
ranking="[1>5,10,13,17,20,21>3>2,4,11,12,18,19>14,16>6,22>7,15>8,9]"
origlen="240" maxlen="72" len="133" maxrate="30.00" rate="55.42"
delta="45.86">
```

ResInterm16. Resultados Intermediários do sumário CIENCIA_2000_6389

Nesse caso, as EDUs 1,5,10,13,17,20 e 21 são consideradas altamente salientes, pois foram incluídas no sumário na exata ordem em que aparecem na classificação de saliência e as EDUs 16 e 19, as quais são S da relação de ATTRIBUTION, devem constar no sumário por determinação interna do sistema. No entanto, esse sumário extrapola em 15% a TC considerada ideal, pois todas as EDUs que apresentam o mesmo peso na classificação de saliência são consideradas.

Ao se calcular a TC a cada EDU saliente incluída, após a inclusão da EDU 17, juntamente com a EDU 16, a qual é S da relação de ATTRIBUTION, o sumário já apresenta uma TC de 33% e deve ser finalizado. O sumário novo é ilustrado em SumNovo6.

[1] A discussão sobre a biotecnologia nacional está enviesada, **[5]** Guerra citou a micropropagação de vegetais **[10]** como exemplo de biotecnologia de baixo custo. [12] Para o agrônomo, **[13]** o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local, como o cultivo de plantas com a capacidade de captar certos elementos presentes na terra. [14] O presidente da Embrapa [16] Alberto Portugal, salientou **[17]** que a empresa busca soluções para os problemas da agricultura nacional.

SumNovo6. Sumário novo do texto CIENCIA_2000_6389

Esse apresenta uma TC mais próxima da ideal e a quebra, introduzida na EDU 5, tanto no sumário novo quanto no antigo, se deve ao *acc*(5)={1,5} não conter a EDU 3, na qual está o antecedente da expressão anafórica e não é decorrente deste proposta, no entanto.

O sumário CIENCIA_2000_17082, ilustrado em Sum19, com respectivos resultados intermediários em ResInterm17 também foi construído somente com EDUs altamente salientes e um S de relação de ATTRIBUTION.

[1] O Instituto Nacional de Pesquisas Espaciais **[3]** prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos, **[5]** Esse é o pior panorama climático previsto pelo instituto, [17] Nobre disse **[18]** 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. **[19]** O Brasil emite 280 milhões de toneladas de carbono **[21]** na atmosfera por ano. **[25]** O desmatamento da Amazônia atingiu 16.926 km2 em 99, **[31]** 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.

Sum19. Sumário do texto CIENCIA_2000_17082

<extract method="rank-prune/marcurank/ranksum"
edus="[1,3,5,17,18,19,21,25,31]" spine="[1,3,18,5,19]" ignored="[]"
ranking="[1,3>18>5,19,21,25,31>4,9,11,17>6,15,23,29,34,36,38>2,7,12,16,22,24
,27,30,32>8,13,26,28,33,35,37,39>14,20>10]" origlen="299" maxlen="89"
len="109" maxrate="30.00" rate="36.45" delta="17.71">

ResInterm17. Resultados Intermediários do sumário CIENCIA_2000_17082

Como a TC é calculada a cada EDU saliente incluída, ao incluir a EDU 25 a TC já atinge os 30% e o sumário novo, constante de SumNovo7, não considera a EDU 31.

[1] O Instituto Nacional de Pesquisas Espaciais **[3]** prediz um aumento de temperatura de até 5C nas áreas mais secas da Amazônia, em 50 anos, **[5]** Esse é o pior panorama climático previsto pelo instituto, [17] Nobre disse **[18]** 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. **[19]** O Brasil emite 280 milhões de toneladas de carbono **[21]** na atmosfera por ano. **[25]** O desmatamento da Amazônia atingiu 16.926 km2 em 99

SumNovo7. Sumário novo do texto CIENCIA_2000_17082

A quebra de clareza referencial, constante da EDU 17, também se deve ao fato de o seu antecedente, presente na EDU 6, não constar no seu *acc*(17)={1,3,17} e não constitui foco deste trabalho.

Outros casos interessantes foram computados durante a análise, porém que não têm relação direta com a TC, como: i) a possibilidade de descarte de EDUs provenientes do *acc* e a inclusão das próximas salientes em sua substituição e ii) a confirmação da necessidade de rebaixamento de EDU saliente em prol da manutenção da TC.

6.6.3 Substituição de EDU proveniente de *acc* por EDU altamente saliente

Observa-se que o sumário do texto CIENCIA_2000_6380, ilustrado em Sum20, não teve nenhuma EDU indicada pelo Modelo de Saliência descartada, pois as primeiras EDUs indicadas em ranking são as que o compõem, como exibido em ResInterm18.

[1] Astrônomos brasileiros esperam fotografar os primeiros planetas fora do Sistema Solar com a ajuda do maior telescópio do mundo, o Gemini, a partir do ano que vem. [6] O planeta mais distante até hoje fotografado é Plutão. **[12]** No entanto, o Gemini poderá "enxergar" planetas. **[19]** O projeto Gemini, resultado de um consórcio de sete países, envolve a construção de dois telescópios com um espelho de oito metros de diâmetro.

Sum20. Sumário do texto CIENCIA_2000_6380

<extract method="rank-prune/marcurank/ranksum" edus="[1,6,12,19]"
spine="[1,19,12,6]"
ignored="[9,17,7,10,22,18,16,8,5,3,15,14,13,11,4,23,21,20,2,24]"
ranking="[1>19>12,24>2,6,20,21,23>4,11,13,14,15>3,5,8,16,18,22>10>7,17>9]"
origlen="250" maxlen="75" len="73" maxrate="30.00" rate="29.20" delta="-2.74">

ResInterm18. Resultados Intermediários do sumário CIENCIA_2000_6380

Os resultados intermediários acima indicam que a EDU 6 foi incluída no sumário por ser proveniente da classificação de saliência; fato evidenciado pela presença das EDUs 2 e 24, grifadas no conjunto ignored, e da sua presença no conjunto spine. Porém, ao verificar os *accs* das EDUs 1, 19 e 12, incluídas pela classificação de saliência, verifica-se que a EDU 6 é proveniente do *acc*(12)={1,6,12} e, portanto, essa não deveria constar em spine. Assim, por ela ser proveniente de *acc*, houve a tentativa de descartá-la e acrescentar a próxima saliente, ou seja, a EDU 24.

Pela aplicação das heurísticas, verificou-se que EDU 6 pôde ser ignorada sem prejuízo algum à mensagem veiculada, como ilustra o sumário constante em SumNovo8. Neste caso, a EDU 24, próxima mais saliente após a escolha das EDUs 1, 12 e 19, é incluída no sumário. Essa redução de *accs* de EDUs privilegia a inclusão de outras EDUs salientes não consideradas no sumário antigo. Isso é evidenciado pela explicação do que é realizado passo a passo a partir do modelo proposto.

[1] Astrônomos <Hprof> brasileiros <Hnat> esperam fotografar os primeiros planetas <Lstar> fora do Sistema Solar <Lstar> com a ajuda do maior telescópio <tool> do mundo, o Gemini <object>, a partir do ano que vem. **[12]** No entanto, o Gemini <object> poderá "enxergar" planetas <Lstar>. **[19]** O projeto Gemini <object>, resultado de um consórcio de sete países, envolve a construção de dois telescópios <tool> com um espelho <tool> de oito metros <unit> de diâmetro <f-q>. [23] Segundo Steiner, **[24]** os telescópios <tool> Gemini <object> têm capacidade científica de observação e qualidade de imagem dez vezes superiores às de um telescópio <tool> espacial.

SumNovo8. Sumário novo do texto CIENCIA_2000_6380

Após a inclusão das EDUs 1 (com *acc*(1)={1}) e 19 (com *acc*(19)={1,19}), inclui-se a EDU 12, com *acc*(12)={1,6,12}. Suas possíveis anáforas são **<object>** e **<Lstar>**. A análise se

aplica somente à EDU 6, já que as demais desse *acc* já foram selecionadas para o sumário novo. A EDU 6 'O planeta **<Lstar>** mais distante até hoje fotografado é Plutão **<Lstar>**' possui duas etiquetas semânticas iguais. Pode-se aplicar a regra A119 (**<object>**,**<Lstar>**), que indica que, para a possível anáfora da EDU 12 marcada com **<object>**, seu possível antecedente é etiquetado com **<Lstar>**. Essa regra, no entanto, não indica elementos textuais correferentes. Porém, comparando-se as etiquetas da EDU 6 e da EDU 1, verifica-se que a EDU 1 apresenta mais possibilidades de antecedentes para as anáforas constantes da EDU 12 e a EDU 6 não é considerada no sumário. Pela aplicação da regra A50 (**<tool>**,**<object>**) aponta-se corretamente para o antecedente da anáfora marcada com **<object>** na EDU 12 e ainda para outro elemento dessa mesma CCR pela coincidência de etiquetas **<object>** nas EDUs 12 e 1. Já a anáfora marcada com **<Lstar>** na EDU 12 tem dois possíveis antecedentes na EDU 1, ambos indicados pela coincidência de etiquetas e um deles é o correto.

A esta altura, o sumário novo em construção contem as EDUs 1, 12 e 19 e como a EDU 12 é a última a ser selecionada (ver conjunto ranking em ResInterm18), a próxima EDU a ser considerada é a EDU 24, lembrando que ela só será considerada porque a TC calculada ainda representa 25% do sumário pretendido. Essa EDU tem como *acc*(24)={1,19,23,24}, sendo que a única candidata a análise é a EDU 23, já que as demais já foram selecionadas. Por estar ligada à EDU 24 através de uma relação de ATTRIBUTION, essa deve ser obrigatoriamente incluída no sumário. Ao incluí-la, o próximo passo é buscar seu *acc*(23)={1,19,23}, o qual não indica novas EDUs a incluir e, assim, a observância das restrições que satisfazem a EDU 24 já está concluída.

Com a inclusão da EDU 23, uma quebra de clareza referencial é introduzida no sumário novo, a qual deixa 'Steiner' sem referente completo. Essa quebra, como as outras mencionadas anteriormente, se deve ao fato do *acc*(23)={1,19,23} e o *acc*(19}={1,19} não conterem a EDU 3, na qual está o antecedente, porém esse não constitui problema proveniente desta proposta.

O sumário novo contém as EDUs 1, 12, 19, 23 e 24 e sua TC é de 35%. Como descreve o Passo 10 do algoritmo apresentado na seção 5.5, levando em consideração a TC do sumário anterior, de 25%, a distância entre os dois é igual, ou seja, o sumário final ultrapassou em 5% a TC ideal e o sumário anterior tinha seu tamanho 5% menor que esse taxa. O sumário novo, então, é formado agora somente por EDUs consideradas altamente salientes. A decisão de manter o sumário antigo ou o sumário novo fica a critério do usuário.

6.6.4 Sumário novo que comprova a necessidade de rebaixamento de saliência

Para um dos sumários antigos que teve a saliência rebaixada, a tentativa de não rebaixamento foi em vão. O sumário do texto CIENCIA_2000_6391, ilustrado em Sum21, exemplifica esse caso. A saliência aqui foi rebaixada, como ilustra as EDUs grifadas no conjunto ranking de ResInterm19.

[1] Sete ativistas do Greenpeace foram presos ontem nos EUA **[3]** Os ativistas **[5]** foram rechaçados pelos tripulantes de um navio de bandeira dinamarquesa [8] O Greenpeace sustenta **[9]** que existe uma grande possibilidade que a madeira seja ilegal. **[13]** A operação resulta de uma investigação de 18 meses sobre o comércio de madeira em regiões remotas da Amazônia.

Sum21. Sumário do texto CIENCIA_2000_6391

```
<extract method="rank-prune/marcurank/ranksum" edus="[1,3,5,8,9,13]"
spine="[1,13,9,3,5,8]" ignored="[6,4,10,14,12,7,2,11,16,15]"
ranking="[1>13>9,15,16>3,5,11>2,7,8,12,14>10>4,6]" origlen="201"
maxlen="60" len="59" maxrate="30.00" rate="29.35" delta="-2.20">
```

ResInterm19. Resultados Intermediários do sumário CIENCIA_2000_6391

Seguindo a classificação de saliência, foram incluídas as EDUs 1,13 e 9, rebaixou-se a classificação de saliência para as EDUs 15 e 16 e incluiu-se as EDUs 3 e 5. Nesse sumário, a EDU 8, única proveniente de *acc*, não pode ser descartada já que é S de uma relação de ATTRIBUTION.

Caso a EDU 15 fosse incluída e fosse possível descartar alguma proveniente do *acc*(15)={1,13,14,15}, o rebaixamento poderia ser evitado, porém a tentativa foi em vão, pois a EDU 13 já consta no sumário e a EDU 14 está ligada à EDU 15 através de uma relação de ATTRIBUTION e não pode ser descartada. O sumário teste resultante ficou com 38% de TC e parte-se para a análise da EDU 16, segunda EDU rebaixada. Seu *acc*(16)={1,13,15,16} indica que além da EDU 16, a 15 deve também ser incluída, que por sua vez, pressupõe a inclusão da 14, pelo motivo acima especificado e esse sumário ficaria com 45% de TC.

A impossibidade da utilização das EDUs rebaixadas no sumário antigo comprova, portanto, a necessidade do rebaixamento da saliência.

6.6.5 Impossibilidade de aplicação do modelo heurístico

O sumário do texto CIENCIA_2000_6381 reflete o caso em que somente EDUs consideradas altamente salientes são consideradas, como se pode observar pela comparação entre Sum22 e ResInterm20.

[1] Após o anúncio do sequenciamento do genoma, na semana passada, a França resiste como único país da União Européia a não permitir patenteamento de genes**.** **[2]** A UE adota, desde junho de 1998, diretiva favorável ao patenteamento de genes. **[8]** A França é o único país que se recusa a aceitar a determinação europeia. **[19]** O assunto deve ser debatido durante a presidência francesa da EU, no segundo semestre.

Sum22. Sumário do texto CIENCIA_2000_6381

```
<extract method="rank-prune/marcurank/ranksum" edus="[1,2,8,19]"
spine="[1,19,2]" ignored="[]"
ranking="[1>19>2,8>3,5,10,11,13>7>6,9,15>4,17>12,14,18>16]" origlen="258"
maxlen="77" len="81" maxrate="30.00" rate="31.40" delta="4.44">
```

ResInterm20. Resultados Intermediários do sumário CIENCIA_2000_6381

Como explicitado na seção 4.2.2, há inconsistências entre os dados constantes de spine e as EDUs presentes no sumário e assume-se que o conjunto spine esteja incompleto nos casos abaixo relacionados. Assim, considera-se que a EDU 8, contante de Sum22, é incluída no sumário por ser proveniente da classificação de saliência e deveria constar de spine.

Como o sumário não contém nenhuma EDU proveniente de *accs*, mas somente EDUs altamente salientes e a TC está próxima da ideal, esta proposta não se aplica a esse caso.

Após a análise de cada caso e com o intuito de comparar efetivamente os sumários aqui gerados com os do VeinSum em termos da preservação da clareza referencial, encontra-se na próxima seção uma síntese detalhada.

6.6.6 Síntese da aplicação das heurísticas para a seleção de unidades textuais correferentes

Dentre os 11 sumários produzidos pelo VeinSum, para 2 deles a proposta original deste trabalho foi aplicada com sucesso, ou seja, pôde-se comprovar a hipótese (iii) de que a aplicação das heurísticas possibilita a redução de *accs*, além da hipótese (vii) de que, com essa redução, o sumarizador passa a respeitar melhor a classificação de saliência das unidades incluídas no sumário. Para os outros 9 casos analisados, as heurísticas foram aplicadas com enfoques distintos, já que nem todos os sumários apresentavam rebaixamento da saliência e resultados interessantes foram obtidos, os quais estão relacionados abaixo:

- o modelo heurístico permite excluir dos *accs* EDUs de importância secundária que seriam escolhidas pelo modelo original e, assim, os sumários novos são construídos com EDUs mais salientes;
- ao considerar a distância dos sumários em construção em relação ao seu tamanho, quando comparado ao tamanho do sumário ideal, o modelo pode garantir sumários cuja TC seja mais próxima da ideal, buscando, assim, satisfazer a hipótese (vi);
- o modelo comprova que o rebaixamento da saliência é realmente necessário em alguns casos, já que ao testar a utilização da EDU que fora descartada pelo VeinSum, verifica-se que a TC é, de fato, corrompida e nenhuma alternativa é encontrada para esses casos;
- a aplicação do modelo permitiu o descarte de EDUs provenientes de *accs* para alguns casos, porém essas EDUs eram as próximas indicadas pela classificação de saliência e tiveram que ser incluídas por esse motivo;
- a aplicação das heurísticas não pôde ser efetuada em sumários que são constituídos somente por EDUs altamente salientes ou provenientes de relações indivisíveis como ATTRIBUTION ou SAME-UNIT, pois essas não podem ser descartadas e tais sumários resultaram iguais aos do VeinSum.

Abaixo, encontram-se quadros que explicitam o panorama geral dos resultados.

Na Tabela 6, a seguir, ilustra-se a comparação dos sumários antigos com os sumários novos com foco no número de EDUs rebaixadas e de EDUs indicadas como altamente salientes incluídas no sumário. São consideradas na coluna ‘EDUs na ordem de saliência’, somente as EDUs incluídas anteriormente a qualquer rebaixamento.

Tabela 6. Quadro de rebaixamento da saliência de EDUs

		Sumários antigos		Sumários novos	
	Texto-Fonte	# EDUs rebaixadas	# EDUs na ordem de saliência	# EDUs rebaixadas	# EDUs na ordem de saliência
1	CIENCIA_2000_6380	0	3	0	4
2	CIENCIA_2000_6381	0	4	0	4
3	CIENCIA_2000_6389	0	7	0	5
4	CIENCIA_2000_6391	2	3	2	3
5	CIENCIA_2000_17082	0	8	0	8
6	CIENCIA_2000_17088	0	4	0	4
7	CIENCIA_2000_17101	15	1	0	2
8	CIENCIA_2000_17108	0	8	0	8
9	CIENCIA_2000_17109	0	4	0	3
10	CIENCIA_2001_19858	3	6	0	8
11	CIENCIA_2000_24212	1	8	1	8
	TOTAIS GERAIS	**21**	**56**	**3**	**57**

Ao observar a média de EDUs rebaixadas, verifica-se que nos sumários novos essas acontecem em número bem menor que nos sumários antigos. O sumário antigo do texto 7, por exemplo, teve 1 EDU da classificação de saliência incluída no sumário, seguida de 15 rebaixamentos não sequenciais, ao passo que o sumário novo desse mesmo texto, teve as duas primeiras EDUs apontadas pela saliência incluídas no sumário e nenhum rebaixamento, ou seja, somente duas EDUs salientes são incluídas e as demais são provenientes dos *accs* das duas primeiras unidades salientes.

Já o sumário antigo do texto 10, teve as 6 primeiras EDUs da classificação de saliência incluídas no sumário, seguidas de 3 rebaixamento sequenciais, sendo que no sumário novo, as 8 primeiras EDUs salientes foram incluídas e não houve nenhum rebaixamento.

Os sumários dos textos 3 e 9, no entanto, apresentam o número de EDUs salientes menor do que os sumários antigos, porém não tiveram nenhuma EDU saliente rebaixada. Vale notar que os sumários antigos, apesar de terem mais EDUs altamente salientes, corrompiam a TC e os sumários novos apresentam essa taxa muito mais próxima da considerada ideal, como se observa no quadro de distâncias entre taxas de compressão exibido na Tabela 7, a seguir.

Tabela 7. Quadro de distância entre taxas de compressão[28]

				Sumários antigos			Sumários novos		
	Texto-Fonte	Tamanho do TF	Tamanho do sum ideal	Tamanho real do sum	TC real do sum (%)	Distância do sum	Tamanho real do sum	TC real do sum (%)	Distância do sum
1	CIENCIA_2000_6380	230	69	67	29	**2**	67	29	**2**
2	CIENCIA_2000_6381	220	66	66	30	**0**	66	30	**0**
3	CIENCIA_2000_6389	220	66	90	41	**24**	73	33	**7**
4	CIENCIA_2000_6391	173	52	53	31	**1,1**	53	31	**1,1**
5	CIENCIA_2000_17082	272	82	106	39	**24,4**	81	30	**0,6**
6	CIENCIA_2000_17088	317	95	124	39	**28,9**	109	34	**13,9**
7	CIENCIA_2000_17101	315	95	84	27	**10,5**	96	30	**1,5**
8	CIENCIA_2000_17108	284	85	90	32	**4,8**	89	31	**3,8**
9	CIENCIA_2000_17109	244	73	101	41	**27,8**	92	38	**18,8**
10	CIENCIA_2001_19858	490	147	155	32	**8**	148	30	**1**
11	CIENCIA_2000_24212	492	148	150	30	**2,4**	150	30	**2,4**
	TOTAIS GERAIS	**3257**	**977**	**1086**	**371**	**134**	**1024**	**347**	**52,1**
	Médias	**296,1**	**88,8**	**98,7**	**33,7**	**12,2**	**93,1**	**31,6**	**4,7**

Verifica-se que a distância dos sumários novos, em relação à distância dos sumários antigos, está mais próxima de 0, que nesse caso representa a TC de 30% considerada. Houve, portanto, uma melhora significativa nesse quesito nos sumários novos, com exceção dos que ficaram iguais aos antigos, os quais apresentam distâncias iguais. De forma geral, ao considerar a média da distância entre os sumários antigos e os novos, isso pode ser comprovado.

A fim de que a TC se aproximasse ao máximo dos 30%, para os casos em que os sumários antigos apresentavam TC menor que a considerada, EDUs foram incluídas no sumário novo e para os sumários antigos que apresentavam TC maior, EDUs provenientes de *accs* foram descartadas.

Já com relação à manutenção da clareza referencial, observa-se na Tabela 8, que o número de quebras de CCRs entre os sumários novos e os antigos permanece praticamente o mesmo para os 11 textos analisados, sendo que, nos sumários antigos, 7 deles não apresentam quebra alguma e 4 apresentam apenas uma quebra, ao passo que, nos sumários novos, 6 não apresentam quebra e 5 deles apresenta apenas uma.

[28] As taxas de compressão apresentadas na Tabela 7 podem não conferir com as apresentadas nas ilustrações de Resultados Intermediários, pois neste trabalho elas são calculadas por número de palavras e no VeinSum o método para chegar a esse cálculo não é explicitado.

Tabela 8. Comparação da clareza referencial dos sumários do VeinSum com os gerados a partir desta proposta

			Sumários antigos			Sumários novos		
	Texto-Fonte	# CCRs no TF	# CCRs no sum	% CCRs do TF no sum	# quebras de CCRs	# CCRs no sum	% CCRs do TF no sum	# quebras de CCRs
1	CIENCIA_2000_6380	10	4	40	0	5	50	1
2	CIENCIA_2000_6381	11	6	55	0	6	55	0
3	CIENCIA_2000_6389	8	6	75	1	6	75	1
4	CIENCIA_2000_6391	6	5	83	1	5	83	1
5	CIENCIA_2000_17082	10	10	100	1	9	90	1
6	CIENCIA_2000_17088	11	7	64	0	7	64	0
7	CIENCIA_2000_17101	17	9	53	0	12	71	0
8	CIENCIA_2000_17108	9	5	56	0	5	56	0
9	CIENCIA_2000_17109	12	10	83	0	10	83	0
10	CIENCIA_2001_19858	12	7	58	1	9	75	1
11	CIENCIA_2000_24212	20	13	65	0	13	65	0
	TOTAIS GERAIS	**126**	**82**	**7,3**	**4**	**87**	**7,7**	**5**
	Médias	**11,5**	**7,5**	**0,7**	**0,4**	**7,9**	**0,7**	**0,5**

Na coluna '# CCRs no sumário' são considerados quaisquer elementos textuais constantes da CCR, mesmo que seja apenas um deles, por exemplo, para uma CCR com 12 elementos constantes do texto-fonte, se o sumário contiver apenas um deles, esse é computado. Apesar de o número de CCRs consideradas nos sumários antigos e os novos serem praticamente iguais, vale ressaltar que as CCRs consideradas podem não ser as mesmas, pois os conteúdos dos sumários podem variar consideravelmente, o que faz variar também as CCRs utilizadas. Observa-se, para os textos 7 e 10 da Tabela 8, que o número de CCRs no sumário novo aumentou em relação ao sumário antigo e poderia se esperar que houvesse mais quebras de clareza referencial, porém isso não ocorreu. Apesar de algumas heurísticas terem levado a antecedentes equivocados em alguns casos, o antecedente da anáfora constava no sumário final na maioria das vezes.

O único sumário novo que apresenta quebra e que essa não ocorre no sumário antigo é o texto 1. Nesse sumário, ocorreu a exclusão de uma EDU proveniente de *acc* e a próxima EDU da classificação de saliência foi incluída em sua substituição, ou seja, o sumário passou a ser constituído somente de EDUs altamente salientes, com a penalidade da quebra de clareza referencial introduzida. Vale lembrar, porém, que as 4 quebras apresentadas nos sumários antigos e as 5 apresentadas nos sumários novos se devem à ausência do termo antecedente de uma mesma expressão anafórica presente tanto no sumário antigo, quanto no sumário novo e ocorrem, pois o termo antecedente não consta no *acc* da unidade que contém a anáfora, o que pode ocorrer devido a falhas no algoritmo da VT ao calcular o *acc* das EDUs ou ainda a

árvores RST mal estruturadas. Essa constatação corrobora com a hipótese (v) de que a aplicação das heurísticas, de forma geral, permite a manutenção da clareza referencial nos sumários, já que se os antecedentes não pertencem ao *acc*, não se atribui o erro a este modelo.

Levando em conta os dados das três tabelas acima, verifica-se que o modelo heurístico proposto prova ser eficiente, ao passo que evita que EDUs altamente salientes sejam descartadas do sumário em prol da manutenção da TC, mantém a clareza referencial nos casos em que antecedentes de expressões anafóricas fazem parte dos *accs* e permite a redução da TC em casos de sumários antigos que a corrompem.

7 CONSIDERAÇÕES FINAIS

A proposta deste trabalho consiste de um modelo heurístico teórico e, assim, não se propõe a sua implementação, dada a natureza linguística da pesquisa. A metodologia utilizada para a correção das etiquetas semânticas e filtragem das regras geradas permitiu delinear relações de similaridade semântica pela ótica linguística, já que as regras foram utilizadas apenas como apoio ao especialista, o que se opõe à ótica estatística proposta por Bick. Sabe-se que este estudo é mais valioso pela metodologia adotada do que pelos resultados em si, pois, além de um *corpus* muito pequeno ter sido considerado, ele serviu tanto para a aprendizagem quanto para a avaliação das regras. Muito embora se tenha tomado o cuidado de gerar as regras de aprendizado com *cross-validation*, esse método é reconhecidamente frágil. Entretanto, não foi possível obter um *corpus* mais significativo, quer para treino, quer para teste, pois seria necessário um esforço humano muito grande. Basta somente notar que houve um esforço significativo para a pós-edição manual da anotação semântica dos textos em uso neste trabalho, sem a qual não seria possível usar as regras geradas pelo Weka para definir as heurísticas.

Além do esforço para a preparação dos dados para a modelagem do conhecimento, seria também necessária a anotação manual das informações de correferência e nova correção da etiquetação semântica de SNs correferentes feita pelo PALAVRAS, a fim de utilizar esses dados como referência na avaliação dos sumários produzidos pela aplicação das heurísticas. Devido a essas limitações, os resultados obtidos não são estatisticamente significantes requerendo um grande aprofundamento no futuro. Este trabalho pôde comprovar que, depois de pós-editada, a anotação semântica do PALAVRAS é interessante para identificar possíveis unidades textuais correferentes e, com isso preservar a clareza referencial dos sumários. Ainda foi possível, com a nova proposta, reduzir o número de rebaixamentos de EDUs salientes, resultando em melhor aproximação da taxa de compressão com a idealmente estipulada. Vale ressaltar que essas observações, no entanto, somente se restringem aos estudos relatados para o *corpus* considerado e caso outro seja utilizado, os resultados podem variar e toda a análise deverá ser refeita, para legitimar a proposta de aprimoramento do VeinSum.

Do ponto de vista linguístico, as principais contribuições deste trabalho são: i) a avaliação crítica do desempenho do VeinSum e detecção de problemas internos do sistema, além do levantamento das principais deficiências apresentadas pelos respectivos sumários; ii)

a avaliação crítica do desempenho do *parser* PALAVRAS e pós-edição manual de sua etiquetagem semântica, a qual compreendeu 41% dos SNs que compõem o *corpus* Summ-it; iii) a detecção de inconsistências de anotação no *corpus* utilizado como referência, principalmente no que tange à anotação RST manual, tanto em termos de segmentação quanto de atribuição inadequada de relações retóricas; iv) a geração de regras de associação e de classificação e posterior filtragem manual, a fim de excluir regras muito específicas ou inadequadas e; v) a proposta do algoritmo que utiliza as regras geradas na identificação de unidades textuais correferentes.

E, finalmente, como continuidade desta pesquisa, destacam-se i) a revisão das árvores RST e da anotação de correferência do *Corpus* Summ-it; ii) a realização de testes mais amplos, dados pela aplicação das heurísticas nos 39 textos restantes do *corpus*; iii) a inversão na ordem de aplicação das heurísticas, ou seja, priorizar a aplicação das regras de classificação, a fim de se avaliar os resultados e julgar qual conjunto deve ser aplicado primeiro, o que pode, eventualmente, culminar no descarte de um dos conjuntos; iv) a investigação de como outros parâmetros de resolução anafórica podem contribuir para a identificação de EDUs mais próximas semanticamente, eventualmente agregando-os ao modelo heurístico aqui proposto; v) a comparação do modelo heurístico com iniciativas de resoluções anafóricas, de fato; vi) a avaliação sistemática dos sumários novos com outras medidas de avaliação já clássicas da área, sendo a principal delas a medida de informatividade, que pode ser obtida com o uso da ferramenta ROUGE (Lin, 2004); vii) a avaliação da inteligibilidade dos sumários novos com a ferramenta Coh-Metrix-Port (Scarton & Aluísio, 2010); viii) a investigação de padrões que identifiquem a possibilidade de exclusão de satélites de relações retóricas sem que haja comprometimento da mensagem pretendida; ix) a verificação se relações de longa distância, conforme discutidas por Wolf & Gibson (2006), não poderiam enriquecer o modelo de seleção de unidades correferentes ou, mesmo, superar a qualidade dos sumários manualmente gerados com base nas heurísticas, a partir de novas formas de estruturação das unidades textuais e, por fim, x) a implementação do modelo heurístico por um cientista da computação, resultando em outra versão do VeinSum, visando confirmar os resultados obtidos nesta pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

Amâncio, M. A. (2009). *Elaboração textual via definição de entidades mencionadas e de perguntas relacionadas aos verbos em textos simplificados do português.* Qualificação de Mestrado. Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo. Junho. São Carlos – SP.

Azzam, S.; Humphreys; K., Gaizauskas, R. (1999). Using correference chains for text summarization. In: *ACL Workshop on Correference and its Applications*. Baltimore.

Beaugrande, R.; Dressler, W. (1981). *Introduction to Textlinguistics*. Londres: Longman.

Bick, E. (2000). *The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a constraint grammar framework*. PhD Thesis. Arhus University, Arhus.

Bick, E. (2005). Gramática Constritiva na análise automática de sintaxe portuguesa. In: *A Língua Portuguesa no Computador*, T. Berber Sardinha (Eds). Campinas – SP: Mercado de Letras.

Bick, E. (2007). Automatic Semantic Role Annotation for Portuguese. In: *Proceedings of the 5th Workshop on Information and Human Language Technology/Anais do XXVII Congresso da SBC*. pp. 1713-1716. Rio de Janeiro – RJ.

Carbonel, T. I. (2007). *Estudo e validação de teorias do domínio linguístico com vistas à melhoria do tratamento de cadeias de correferência em sumarização automática*. Dissertação de Mestrado. Departamento de Letras. Universidade Federal de São Carlos. Agosto. São Carlos – SP.

Chaves, A. R. (2007). *A resolução de anáforas pronominais da língua portuguesa com base no algoritmo de Mitkov.* Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. Agosto. São Carlos - SP.

Coelho, J. C. B.; Muller, V. M.; Abreu, S. C.; Vieira, R.; Rino, L. H. M. (2006). Resolving Potuguese Nominal Anaphora. In: *7th Workshop on Computational Processing of Portuguese Language.* Vol. 3960. pp. 160-169. Itatiaia: Springer.

Collovini, S.; Carbonel, T. I.; Fuchs, J. T.; Coelho, J. C.; Rino, L. H. M.; Vieira, R. (2007). Summ-it: Um *corpus* anotado com informações discursivas visando à Sumarização Automática. In: *Proceedings of the V Workshop on Information and Human Language Technology*, V. Quental & C. Oliveira (Eds). XXVII Congresso da Sociedade Brasileira de Computação. Rio de Janeiro - RJ.

Collovini, S.; Ribeiro-Junior, L. C.; Gonçalves, P. N.; Muller, V.; Vieira, R. (2008). Using semantic prototypes for discourse status classification. In: *Computational Processing of the Portuguese Language*. Lecture Notes in Computer Science. Vol. 5190. pp. 236-239: Springer Berlin.

Cristea, D.; Ide, N.; Romary, L. (1998). Veins Theory: a model of global discourse cohesion and coherence. In: *Proceedings of the Coling/ACL*. pp. 281-285. Montreal, Canadá.

Cristea, D. (2003). The relationship between discourse structure and referentiality in Veins Theory. In: *Natural Language Processing between Linguistic Inquiry and System Engineering*, W. Menzel and C. Vertan (Eds.). Iasi: University Publishing House.

Cristea, D.; Postolache, O.; Pistol, I. (2005). Summarization through discourse structure. In: *Proceedings of the 6th International Conference on Computational Linguistics and Intelligence Text Processing*, A. Gebukh (Ed). Mexico City, Mexico.

Cristea, D. (2005). Motivations and Implications of Veins Theory. In: B. Sharp (Ed.) *Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science*. pp. 32-44. Miami, U.S.A.

Dias-da-Silva, B. C.; Montilha, G.; Rino, L. H. M.; Specia, L.; Nunes, M. G. V.; Oliveira Jr., O. N.; Martins, R. T.; Pardo, T. A. S. (2007). *Introdução ao Processamento das Línguas Naturais e Algumas Aplicações*. Série de Relatórios Técnicos do NILC. NILC-TR-10-07.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambrigde, MA: The MIT Press.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. (2009). The WEKA Data Mining Software: an update. In: *SIGKDD Explorations*. Vol. 11, Issue 1.

Halliday, M. A. K.; Hasan, R. (1976). *Cohesion in English*. London: Longman.

Jorge, M. L. C. (2010). *Sumarização Automática Multidocumento: seleção de conteúdo com base no Modelo CST (Cross-document Structure Theory)*. Dissertação de Mestrado.

Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo. Fevereiro. São Carlos – SP.

Jorge, M. L. C.; PARDO, T. A. S. (2010). Experiments with CST-based Multidocument Summarization. In: *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing.* pp. 74-82. Uppsala.

Koch, I. G. V. (2004a). *Introdução à Linguística Textual: trajetória e grandes temas*. São Paulo: Martins Fontes.

Koch, I. G. V. (2004b). *A coesão textual*. São Paulo: Contexto.

Koch, I. G. V.; Travaglia, L. C. (2004). *A coerência textual*. São Paulo: Contexto.

Lin, C. (2004). ROUGE: a package for automatic evaluation of summaries. In: *Proceedings of the Workshop on Text Summarization Branches Out*. pp. 74-81.

Mani, I.; Maybury, M. T. (1999). *Advances in Automatic Text Summarization*. Cambridge, MA: The MIT Press.

Mani, I. (2001). *Automatic Summarization*. Amsterdan: John Benjamin's Publishing Co.

Mann, W. C.; Thompson, S. A. (1988). *Rhetorical structure theory: a theory of text organization.* Technical Report ISI/RS-87-190.

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

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

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

Marcuschi, L. A. (1983). *Linguística de texto: como é e o que se faz*. Recife: Universidade Federal de Pernambuco, Série Debates 1.

Mitkov, R. (2002). *Anaphora Resolution*. London: Pearson ESL.

O'Donnell, M. (1997). RSTTool: an RST analysis tool. In: *Proceedings of the 6th European Workshop on Natural Language Generation*. Duisburg, Germany: Gerhard-Mercator University.

Ono, K.; Sumita, K.; Miike, S. (1994). Abstract generation based on rhetorical structure extraction. In: *Proceedings of the 15th International Conference on Computational Linguistics.* Kyoto, Japan.

Pardo, T. A. S. (2002). *DMSumm: um gerador automático de sumários*. Dissertação de Mestrado. Departamento de Computação. Universidade Federal de São Carlos. Abril. São Carlos – SP.

Pardo, T. A. S. (2005). *Métodos para análise discursiva automática*. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo. Junho. São Carlos - SP.

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

Scarton, C. E.; Aluísio, S. M. (2010). Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. In: *Revista Linguamática (Revista para o Processamento Automático das Línguas Ibéricas)* 2(1). pp. 45-61.

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. Universidade Federal de São Carlos. Agosto. São Carlos – SP.

Souza, J. G. C.; Gonçalves, P. G.; Vieira, R. (2008). Learning Coreference Resolution for Portuguese Texts. In: *Computational Processing of the Portuguese Language.* Lecture Notes in Computer Science. Vol. 5190. pp. 153-162. Springer Berlin.

Sparck Jones, K. (1993). What might be in a summary? In: *Information Retrieval 93*, G. Knorz, J. Krause and C. Womser-Hacker (Eds). pp. 9-23. Konstanz: Universitatsverlag Konstanz.

Tomazela, E. C.; Rino, L. H. M. (2009). O uso de informações semânticas do tratar a informatividade de sumários automáticos com foco na clareza referencial. In: *Anais do VII Encontro Nacional de Inteligência Artificial*, A. Villavicencio (Ed.). pp. 799-808. XXIX Congresso da Sociedade Brasileira de Computação (CSBC 2009). Bento Gonçalves – RS.

Tomazela, E. C.; Rino, L. H. M. (2010). *Correção da etiquetagem semântica do Parser PALAVRAS para o Corpus Summ-it*. Série de Relatórios do NILC. NILC – TR-02-10.

Uzêda, V. R.; Pardo, T. A. S.; Nunes, M. G. V (2009). A comprehensive summary informativeness evaluation for RST-based summarization methods. In: *International Journal of Computer Information Systems and Industrial Management Application.*, Vol. 1, pp. 188-196.

Uzêda, V. R.; Pardo, T. A. S.; Nunes, M. G. V (2010). A comprehensive comparative evaluation of RST-based summarization methods. In: *ACM Transactions on Speech and Language Processing*. Vol. 6, pp. 1-20.

Vieira, R. (1998). *Definite description processing in unrestricted text*. PhD Thesis. Computing Department. University of Edinburgh. Edinburgh.

Wolf. F.; Gibson E. (2006). *Coherence in Natural Language: data structures and applications.* Cambridge, MA: The MIT Press.

Apêndice A. Textos-Fonte do Córpus Summ-it analisados

CIENCIA_2000_6380
Astrônomos brasileiros esperam fotografar os primeiros planetas fora do Sistema Solar com a ajuda do maior telescópio do mundo, o Gemini, a partir do ano que vem. A informação foi dada na conferência "Da Origem do Universo aos Grandes Telescópios", ministrada ontem por João Steiner, astrofísico da USP, durante a 52ª Reunião Anual da Sociedade Brasileira para o Progresso da Ciência. A reunião começou anteontem no campus da UnB (Universidade de Brasília). O planeta mais distante até hoje fotografado é Plutão. "Imagine alguém de fora da nossa galáxia que tente fotografar o Sol. Essa pessoa não perceberia a Terra, que seria ofuscada pela luminosidade solar", explicou Steiner. O mesmo acontece atualmente com os astrônomos que fotografam estrelas muito distantes. No entanto, o Gemini poderá "enxergar" planetas. Isso porque, além de medir luz, o telescópio mede calor, por meio de radiação infravermelha. E, embora não emitam luz, planetas geram calor. O projeto Gemini, resultado de um consórcio de sete países, envolve a construção de dois telescópios com um espelho de oito metros de diâmetro. Um dos telescópios já está pronto e em funcionamento no Havaí, EUA. O outro entrará em operação no ano que vem, em Atacama, Chile, e o Brasil terá direito a usá-lo 14 noites por ano. Segundo Steiner, os telescópios Gemini têm capacidade científica de observação e qualidade de imagem dez vezes superiores às de um telescópio espacial.
No. de palavras: 230
CIENCIA_2000_6381
Após o anúncio do sequenciamento do genoma, na semana passada, a França resiste como único país da União Europeia a não permitir patenteamento de genes. A UE adota, desde junho de 1998, diretiva favorável ao patenteamento de genes. O texto, redigido pelo Parlamento Europeu, Comissão Europeia e Conselho de Ministros, utiliza o princípio de que "o genoma não é patenteável, mas a sequência de um gene pode ser". No entanto, há restrições. O patenteamento só pode ser aplicado em pesquisas ligadas a doenças genéticas em que o funcionamento do gene é detalhado. A França é o único país que se recusa a aceitar a determinação europeia. A ministra da Justiça do país, Elisabeth Guigou, disse que a norma é incompatível com as leis francesas de bioética. No início do mês, o CCNE (Comitê Consultivo Nacional de Ética), órgão que orienta o governo francês sobre aspectos éticos da biotecnologia, reforçou a posição da ministra, alegando que "o conhecimento da sequência de um gene não pode ser assimilado como produto patenteado e, portanto, não é patenteável". "Bem comum da humanidade, (o sequenciamento de genes) não pode ser limitado por patentes que pretendem, em nome do direito de propriedade industrial, proteger a exclusividade desse conhecimento", diz parecer do CCNE. O assunto deve ser debatido durante a presidência francesa da UE, no segundo semestre.
No. de palavras: 220
CIENCIA_2000_6389
A discussão sobre a biotecnologia nacional está enviesada, pois está sendo entendida

como sinônimo de transgenia. A opinião é do agrônomo Miguel Guerra, da UFSC (Universidade Federal de Santa Catarina).
Guerra citou a micropropagação de vegetais (produção de mudas em laboratório, feita para evitar doenças e selecionar vegetais saudáveis) como exemplo de biotecnologia de baixo custo. Com ela, aumentou-se a produção de moranguinho, no sul do país, de 3,2 kg para 60 kg por hectare.
Para o agrônomo, o Brasil deve buscar o desenvolvimento de transgenias que tentem melhorar as condições da agricultura local, como o cultivo de plantas com a capacidade de captar certos elementos presentes na terra.
O presidente da Embrapa (Empresa Brasileira de Pesquisa Agropecuária), Alberto Portugal, salientou que a empresa busca soluções para os problemas da agricultura nacional. Ele citou o exemplo de pesquisas que, por meio de engenharia genética, buscam obter mamão livre de vírus e feijão também resistente a vírus, culturas de interesse para exportação e consumo interno.
Portugal disse que os agronegócios correspondem a 25% do PIB brasileiro e que a biotecnologia é fundamental para manter a competitividade da agricultura.

No. de palavras: 187

CIENCIA_2000_6391

Sete ativistas do Greenpeace foram presos ontem nos EUA ao tentar impedir o descarregamento de madeira brasileira na cidade de Savannah, na costa do Estado da Geórgia.
Os ativistas (norte-americanos e europeus) foram rechaçados pelos tripulantes de um navio de bandeira dinamarquesa e presos pela polícia.
O navio trazia compensados de madeira exportados pela Selvaplac, subsidiária brasileira de um grupo da Malásia.
O Greenpeace sustenta que existe uma grande possibilidade que a madeira seja ilegal.
"Não temos certeza de que aquela carga era ilegal, mas sabemos que 80% da atividade madeireira no Brasil é irregular e que a Selvaplac tem uma tradição de envolvimento com madeira ilegalmente extraída", disse à Folha Rebeca Lerer, ativista brasileira do Greenpeace.
A operação resulta de uma investigação de 18 meses sobre o comércio de madeira em regiões remotas da Amazônia.
Para a ONG, há evidências de que as companhias que mais exportam madeira para os EUA estejam envolvidas com o comércio ilegal do produto.
A União estima que 80% da madeira extraída na Amazônia brasileira seja ilegal.

No. de palavras: 173

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: 317

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_2001_19858

Cientistas do Centro de Estudos de Saclay, na França, parecem ter encontrado o primeiro indício de que a Via Láctea esteja cercada por um verdadeiro campo minado de buracos negros _cerca de 1 milhão ao todo, eles estimam.
A pesquisa, que sai publicada hoje na revista britânica "Nature" (www.nature.com), é a primeira a constatar que um buraco negro está orbitando o centro da Via Láctea fora do plano do disco galáctico. Atualmente, o objeto está a aproximadamente 6.000 anos-luz do Sol _uma distância pequena, se comparada ao tamanho da galáxia, com seus 100 mil anos-luz de diâmetro.
O objeto em questão _um sistema duplo composto por um buraco negro sendo orbitado por uma estrela anã_ já havia sido detectado desde o ano passado, mas só agora os cientistas puderam se assegurar de que ele não pertencia ao disco galáctico.
Um buraco negro é o resultado do colapso de uma estrela muito maciça. A força gravitacional existente em suas imediações é tão forte que nem a luz consegue escapar. Por isso é tão complicado identificar esses objetos. Atualmente, o único meio seguro de aferir sua presença é observar astros próximos que estejam sob influência de sua gravidade.
A equipe rastreou a trajetória do buraco negro usando novas observações e a análise de dados anteriores, coletados por 43 anos.
A partir dessas observações, eles calcularam a órbita do astro e constataram que se trata de um objeto que passa a maior parte do tempo no halo galáctico _a região que circunda a Via Láctea.
Além de "morar" no halo, aparentemente ele também "nasceu" lá, dizem os cientistas. "Há duas boas evidências de que ele tenha se formado no próprio halo", disse à Folha Felix Mirabel, pesquisador que liderou o grupo. "Uma é a trajetória em si, outra é o quanto a estrela vizinha foi consumida pelo buraco negro, o que indica que o conjunto é

muito velho."
Um buraco negro nascido na própria Via Láctea que tivesse sido atirado para fora posteriormente, provavelmente pela explosão de uma estrela (fenômeno conhecido como supernova), era justamente o que Mirabel vinha procurando desde que iniciou sua pesquisa, há cinco anos. Ainda não foi desta vez que ele achou, mas a meta pode não estar longe.
Seu grupo está atualmente concluindo um segundo estudo, que deve nos próximos meses ser submetido a uma revista científica, abordando não um, mas cinco astros, sendo um deles o mencionado na pesquisa publicada hoje.
O material incluirá um segundo buraco negro pertencente ao halo galáctico. E um dos astros "parece ter um sinal de que tenha sido 'chutado' para fora da galáxia", diz Irapuan Rodrigues, pesquisador brasileiro do grupo, que pela primeira vez publica na "Nature".
A nova pesquisa promete ter impacto ainda maior no meio científico do que o estudo atual. Segundo Rodrigues, a equipe vai propor uma nova classificação para objetos do tipo, dividindo-os entre os pertencentes ao halo e os que vêm do disco galáctico.

No. de palavras: 490

CIENCIA_2003_24212

Biotecnólogos e ambientalistas travaram uma rara aliança na semana passada para comemorar um episódio singular: duas vacas deram cria em Iowa, EUA. Os filhotes não são delas, mas clones de outra espécie, o banteng, um tipo de gado ameaçado de extinção.
A realização é fruto de uma parceria entre a companhia de biotecnologia Advanced Cell Technology, de Massachusetts, o Centro Sioux, de Iowa, e o Centro para Reprodução de Espécies Ameaçadas da Sociedade Zoológica de San Diego, na Califórnia.
Os dois nascimentos, ocorridos em 1º e 3 de abril, marcam o início de uma nova fase para um projeto que já atraía interesse _o Frozen Zoo (ou Zoológico Congelado, na tradução para o português).
A ideia, iniciada em 1976, era coletar e preservar criogenicamente (em baixas temperaturas) amostras celulares de animais ameaçados de extinção, com a esperança de estudá-los e, quem sabe, ressuscitá-los quando a tecnologia assim o permitisse.
Hoje, a coleção do Frozen Zoo, que é coordenado pelo geneticista Oliver Ryder, é a maior do mundo. São cerca de 1.800 amostras e 335 espécies diferentes com o material genético preservado.
O material vai desde os notórios pandas e condores até os menos conhecidos bantengs _parentes asiáticos raros do gado comum que estão à beira do esquecimento. Hoje há menos de 10 mil exemplares remanescentes.
A primeira tentativa de trazer um membro do Frozen Zoo de volta do mundo dos animais perdidos foi com um gauro, outra espécie rara de gado. Uma gravidez acabou levando ao nascimento de um animal, em 2001. O clone morreu dois dias depois.
Os dois bantengs produzidos em Iowa já viveram mais do que isso. Um deles, nascido no dia 1º, está em boa saúde. O segundo está mais debilitado. Mas Ryder não arrisca palpite sobre qual deles pode sobreviver. "Até mesmo o que está bem pode mudar de condição da noite para o dia", diz.
A fase crítica para os dois animais é de duas a três semanas. Depois desse período, quem sobreviver será levado ao Zôo de San Diego, onde viverá com seus colegas de espécie. A equipe não tem planos futuros de clonagem no momento. "Agora queremos ver como esses animais se comportam e como se reproduzem", diz Ryder. "Esse é um projeto para os próximos cinco ou seis anos."
Os dois filhotes, que ainda não têm nome, são cópias genéticas idênticas de um banteng

macho que morreu no Parque Selvagem Animal de San Diego em 1980. No início, a ACT preparou 30 óvulos de vaca, extraindo-lhes o núcleo e fundindo-os com células adultas preservadas do banteng de San Diego. A fusão, feita com ajuda de um choque elétrico, faz com que o óvulo se comporte como se tivesse sido fertilizado, iniciando o processo de divisão celular e criando um embrião. Ele é então implantado numa vaca, que serve como barriga de aluguel. Clonar animais continua sendo uma tarefa difícil. Dos 30 embriões originais, 11 resultaram em gravidez. Desses, só dois chegaram ao nascimento.
No. de palavras: 492

Apêndice B. Elenco de protótipos semânticos providos pelo PALAVRAS

Semantic tags for nouns[29]

PALAVRAS assigns angle-bracketed semantical tags for most nouns and verbs and some adjectives. The 157 semantic tags used for nouns are prototype classes, like <Hprof> for 'professional', which again translate into a subset of atomic features taken from a list of 16 values. The semantic tags are bilingually motivated (Portuguese-Danish translation alternatives) and polysemic words will thus have several tags. The semantical subsystem is in an experimental stage, and not subject to a full disambiguation at the present time, though it can - together with the valency subsystem - yield a fair degree of polysemy resolution even now. The noun tag list below is in alphabetical order, with uppercase tags first.

Animal prototypes:

<A> Animal, umbrella tag (*clone, fêmea, fóssil, parasito, predador*)
<AA> Group of animals (*cardume, enxame, passarada, ninhada*)
<Adom> Domestic animal or big mammal (likely to have female forms etc.: *terneiro, leão/leoa, cachorro*)
<AAdom> Group of domestic animals (*boiada*)
<Aich> Water-animal (*tubarão, delfim*)
<Amyth> Mythological animal (*basilisco*)
<Azo> Land-animal (*raposa*)
<Aorn> Bird (*águia, bem-te-vi*)
<Aent> Insect (*borboleta*)
<Acell> Cell-animal (bacteria, blood cells: *linfócito*)

Plant prototypes:

<B> Plant, umbrella tag
<BB> Group of plants, plantation (field, forest etc.: *mata, nabal*)
<Btree> Tree (*oliveira, palmeira*)
<Bflo> Flower (*rosa, taraxaco*)
<Bbush> Bush, shrub (*rododendro, tamariz*)

cp. also <fruit> (fruit, berries, nuts: *maçã, morango, avelã, melancia*)
further proposed categories: <Bveg> (vegetable *espargo, funcho*)

[29] Retirado do sítio: http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html#semtags em 21/04/2009

Human prototypes:

<H> Human, umbrella tag
<HH> Group of humans (organisations, teams, companies, e.g. *editora*)
<Hattr> Attributive human umbrella tag (many *-ista, -ante*)
<Hbio> Human classified by biological criteria (race, age etc., *caboclo, mestiço, bebé, adulto*) **<Hfam>** Human with family or other private relation (*pai*, *noiva*)
<Hideo> Ideological human (*comunista*, implies <Hattr>), also: follower, disciple (*dadaista*)
<Hmyth> Humanoid mythical (gods, fairy tale humanoids, *curupira, duende*)
<Hnat> Nationality human (*brasileiro*, *alemão*), also: inhabitant (*lisboeta*)
<Hprof> Professional human (*marinheiro*, implies <Hattr>), also: sport, hobby (*alpinista*)
<Hsick> Sick human (few: *asmático, diabético*, cp <sick>)
<Htit> Title noun (*rei*, *senhora*)

Place and spatial prototypes:

<L> Place, umbrella tag
<Labs> Abstract place (*anverso. auge*)
<Lciv> Civitas, town, country, county (equals <L> + <HH>, *cidade, país*)
<Lcover> Cover, lid (*colcha, lona, tampa*)
<Lh> Functional place, human built or human-used (*aeroporto, anfiteatro*, cp. <build> for just a building)
<Lopening> opening, hole (*apertura, fossa*
<Lpath> Path (road, street etc.: *rua, pista*)
<Lstar> Star object (planets, comets: *planeta, quasar*)
<Lsurf> surface (*face, verniz*, cp. <Lcover>)
<Ltip> tip place, edge (*pico, pontinha*, cp. <Labs>)
<Ltop> Geographical, natural place (*promontório, pântano*)
<Ltrap> trap place (*armadilha, armazelo*)
<Lwater> Water place (river, lake, sea: *fonte, foz, lagoa*)

cp. also <bar> (barrier), <build> (building), <inst> (institution), <pict> (picture), <sit> (situation)
cp. also **position prototypes:** <pos-an> (anatomical position), <pos-soc> (social position)

Vehicle prototypes:

<V> Vehicle, umbrella tag and ground vehicle (car, train: *carro, comboio, tanque, teleférico*)
<VV> Group of vehicles (armada, convoy: *frota, esquadra*)
<Vwater> Water vehicle (ship: *navio, submersível, canoa*)
<Vair> Air vehicle (plane: *hidroplano, jatinho*)

Abstract prototypes:

<ac> Abstract countable, umbrella tag (*alternativa, chance, lazer*)
<ac-cat> Category word (*latinismo, número atômico*)
<ac-sign> sign, symbol (*parêntese, semicolcheia*)
<am> Abstract mass/non-countable, umbrella tag (still contains many cases that could be <f-..>, e.g. *habilidade, legalidade*)
<ax> Abstract/concept, neither countable nor mass (*endogamia*), cp. <f>, <sit> etc.

cf. also <f...> (features), <dir> (direction), <geom...> (shapes), <meta> ("transparent" noun)
cf. also **concept prototypes:** <conv> (convention), <domain>, <ism> (idelology), <genre>, <ling> (language), <disease>, <state...>, <therapy>
cf. also **quantity prototypes:** <unit>, <amount>, <cur> (currency), <mon> (money amount)

Action prototypes:

<act> Action, umbrella tag (+CONTROL, PERFECTIVE)
<act-beat> beat-action (thrashing, *pancada, surra*)
<act-d> do-action (typically dar/fazer + N, *tentativa, teste, homenagem*)
<act-s> speech act or communicative act (*proposta*, *ordem*)
<act-trick> trick-action (cheat, fraud, ruse, *jeito, fraude*, similar to <act-d>)
<activity> Activity, umbrella tag (+CONTROL, IMPERFECTIVE, *correria, manejo*)

cp. also <fight>, <dance>, <sport>, <game>, <therapy>

Anatomical prototypes:

<an> Anatomical noun, umbrella tag (*carótida, clítoris, dorso*)
<anmov> Movable anatomy (arm, leg, *braço, bíceps, cotovelo*)
<anorg> Organ (heart, liver, *hipófise, coração, testículo*)
<anost> Bone (*calcâneo, fíbula, vértebra*)
<anzo> Animal anatomy (*rúmen, carapaça, chifres, tromba*)
<anorn> Bird anatomy (*bico, pluma*)
<anich> Fish anatomy (few: *bránquias, siba*)
<anent> Insect anatomy (few: *tentáculo, olho composto*)
<anbo> Plant anatomy (*bulbo, caule, folha*)

cp. also <f-an> (human anatomical feature)

<amount> quantity noun (*bocada, teor, sem-fim*)
<bar> barrier noun (*dique, limite, muralha*)
<build> building (*casa, citadela, garagem*)

Thing prototypes:

<cc> Concrete countable object, umbrella tag (*briquete, coágulo*, normally movable things, unlike <part-build>)
<cc-h> Artifact, umbrella tag (so far empty category in PALAVRAS)
<cc-beauty> ornamental object (few: *guirlanda, rufo*)
<cc-board> flat long object (few: board, plank, *lousa, tabla*)
<cc-fire> fire object (bonfire, spark, *chispa, fogo, girândola*)
<cc-handle> handle (*garra, ansa, chupadouro*)
<cc-light> light artifact (*lampião, farol, projector*)
<cc-particle> (atomic) particle (few: *cátion, eletrônio*)
<cc-r> read object (*carteira, cupom, bilhete, carta*, cf. <sem-r>)
<cc-rag> cloth object (towel, napkin, carpet, rag) , cp. <mat-cloth>
<cc-stone> (= cc-round) stones and stone-sized round objects (*pedra, itá, amonite, tijolo*)
<cc-stick> stick object (long and thin, *vara, lançe, paulito*)

cp. also <con> (container), <cord> (cord), <furn> (funiture), <pict> (picture), <tube>, <clo...> (clothing), <tool...>

Substance prototypes:

<cm> concrete mass/non-countable, umbrella tag, substance (cf. <mat>, *terra, choça, magma*)
<cm-h> human-made substance (cf. <mat>, *cemento*)
<cm-chem> chemical substance, also biological (*acetileno, amônio, anilina, bilirrubina*
<cm-gas> gas substance (so far few: *argônio*, overlap with. <cm-chem> and <cm>)
<cm-liq> liquid substance (*azeite, gasolina, plasma*, overlap with <food> and <cm-rem>)
<cm-rem> remedy (medical or hygiene, *antibiótico, canabis, quinina*, part of <cm-h>, overlap with <cm-chem>)

cp. also <mat...> (materials)

Clothing prototypes:

<cloA> animal clothing (*sela, xabraque*)
<cloH> human clothing (*albornoz, anoraque, babadouro, bermudas*)
<cloH-beauty> beauty clothing (e.g. jewelry, *diadema, pendente, pulseira*)
<cloH-hat> hat (*sombrero, mitra, coroa*)
<cloH-shoe> shoe (*bota, chinela, patim*)

Collective prototypes:

<coll> set, collective (random or systematic collection/compound/multitude of similar but distinct small parts, *conjunto, série*)
<coll-cc> thing collective, pile (*baralho, lanço*)
<coll-B> plant-part collective (*buquê, folhagem*)
<coll-sem> semantic collective, collection (*arquivo, repertório*)
<coll-tool> tool collective, set (*intrumentário, prataria*)

cp. also <HH> (group), <AA> (herd), <BB> (plantation), <VV> (convoy)

<col> coulor (*amarelo, carmesim, verde-mar*)
<con> container (implies <num+> quantifying, *ampola, chícara, aquário*)
<conv> convention (social rule or law, *lei, preceito*)
<cord> cord, string, rope, tape (previously <tool-tie>, *arame, fio, fibrila*)
<cur> currency noun (countable, implies <unit>, cf. <mon>, *dirham, euro, real, dólar*)
<dance> dance (both <activity>, <genre> and <sem-l>, *calipso, flamenco, forró*)
<dir> direction noun (*estibordo, contrasenso, norte*)
<domain> domain (subject matter, profession, cf. <genre>, *anatomia, citricultura, dactilografia*)
<drink> drink (*cachaça, leite, guaraná, moca*)

Time and event prototypes:

<dur> duration noun (test: *durar+*, implies <unit>, e.g. *átimo, mês, hora*)
cf. <per> (period) and <temp> (time point)
<event> event (-CONTROL, PERFECTIVE, *milagre, morte*)
cp. also <occ> (organized event), <process>, <act...> and <activity>

Feature prototypes:

<f> feature/property, umbrella tag (*problematicidade, proporcionalidade*)
<f-an> anatomical "local" feature, includes countables, e.g. *barbela, olheiras*)
<f-c> general countable feature (*vestígio, laivos, vinco)*
<f-h> human physical feature, not countable (*lindura, compleição*, same as <f-phys-h>, cp. anatomical local features <f-an>)
<f-psych> human psychological feature (*passionalidade, pavonice*, cp. passing states <state-h>)
<f-q> quantifiable feature (e.g. *circunferência, calor*, DanGram's <f-phys> covers both <f> and <f-q>)
<f-right> human social feature (right or duty): e.g. *copyright, privilégio, imperativo legal*)
cp. also **state prototypes:** <state>, <state-h> (human state)

Food prototypes:

> **<food>** natural/simplex food (*aveia, açúcar, carne*, so far including <spice>)
> **<food-c>** countable food (few: *ovo, dente de alho*, most are <fruit> or <food-c-h>)
> **<food-h>** human-prepared/complex culinary food (*caldo verde, lasanha*)
> **<food-c-h>** culinary countable food (*biscoito, enchido, panetone, pastel*)
>
> cp. also <drink>, <fruit>
> further proposed categories: <spice>

<fight> fight, conflict (also <activity> and +TEMP, *briga, querela*)
<fruit> fruit, berry, nut (still mostly marked as <food-c>, *abricote, amora, avelã, cebola*)
<furn> furniture (*cama, cadeira, tambo, quadro*)

Concept prototypes:

> **<game>** play, game (*bilhar, ioiô, poker*, also <activity>)
> **<genre>** genre (especially art genre, cf. <domain>, *modernismo, tropicalismo*)
>
> cp. also <conv> (convention), <dance>, <domain>, <ism> (idelology), <ling> (language), <disease>, <sport>, <state...>, <therapy>

<geom> geometry noun (circle, shape, e.g. *losango, octógono, elipse*)
<geom-line> line (few: *linha, percentil, curvas isobáricas*)

<inst> "institution", functional structure (+PLACE, +HUM, *auto-escola, bolsa, cinemateca*) , cp. <Lh> (human-made place) and <HH> (group, organisation)
<ism> ideology or other value system (*anarquismo, anti-ocidentalismo, apartheid*)
<ling> language (*alemão, catalão, bengali*)
<mach> machine (complex, usually with moving parts, *betoneira, embrulhador, limpa-pratos*, cp. <tool>)

<mat> material (*argila, bronze, granito*, cf. <cm>)
<mat-cloth> cloth material (*seda, couro, vison, kevlar*), cp. <cc-rag>

<meta> meta noun (*tipo*, *espécie*)
<mon> amount of money (*bolsa*, *custo*, *imposto*, cf. <cur>)
<month> month noun/name (*agosto, julho*, part of <temp>)
<occ> occasion, human/social event (*copa do mundo, aniversário, jantar, desfile*, cp. unorganized <event>)
<per> period of time (prototypical test: *durante*, e.g. *guerra, década*, cf. <dur> and <temp>)

Part prototypes:

<part> distinctive or functional part (*ingrediente, parte, trecho*)
<part-build> structural part of building or vehicle (*balustrada, porta, estai*)
<piece> indistinctive (little) piece (*pedaço, raspa*)

cf. other structurals, such as <cc-handle>, <Ltip>

Perception prototypes:

<percep-f> what you feel (senses or sentiment, pain, e.g. *arrepio, aversão, desagrado, cócegas*, some overlap with <state-h>)
<percep-l> sound (what you hear, *apitadela, barrulho, berro, crepitação*)
<percep-o> olfactory impression (what you smell, *bafo, chamuscom fragrância*)
<percep-t> what you taste (PALAVRAS: not implemented)
<percep-w> visual impression (what you see, *arco-iris, réstia, vislumbre*)

<pict> picture (combination of <cc>, <sem-w> and <L>, *caricatura, cintilograma, diapositivo*)

<pos-an> anatomical/body position (few: *desaprumo*)
<pos-soc> social position, job (*emprego, condado, capitania, presidência*)
<process> process (-CONTROL, -PERFECTIVE, cp. <event>, *balcanização, convecção, estagnação*)

Semantic product prototypes:

<sem> semiotic artifact, work of art, umbrella tag (all specified in PALAVRAS)
<sem-c> cognition product (concept, plan, system, *conjetura, esquema, plano, prejuízo*)
<sem-l> listen-work (music, *cantarola, prelúdio*, at the same time <genre>: *bossa nova*)
<sem-nons> nonsense, rubbish (implies <sem-s>, *galimatias, farelório*)
<sem-r> read-work (*biografia, dissertação, e-mail, ficha cadastral*)
<sem-s> speak-work (*palestra, piada, exposto*)
<sem-w> watch-work (*filme, esquete, mininovela*)

cp. <ac-s> (speach act), <talk>
cf. also **concept prototypes:** <conv> (convention), <domain>, <ism> (idelology), <game>, <genre>, <ling> (language), <disease>, <state...>, <therapy>

<sick> disease (*acne, AIDS, sida, alcoolismo*, cp. <Hsick>)
<sick-c> countable disease-object (*abscesso, berruga, cicatriz, gangrena*)

State-of-affairs prototypes:

<sit> psychological situation or physical state of affairs (*reclusão, arruaça, ilegalidade*, more complex and more "locative" than <state> and <state-h>
<state> state (of something, otherwise <sit>), *abundância, calma, baixa-mar, equilíbrio*
<state-h> human state (*desamparo, desesperança, dormência, euforia, febre*, cp. <f-psych> and <f-phys-h>, which cover innate features)

<sport> sport (*capoeira, futebol, golfe*, also <activity> and <domain>)

<talk> speech situation, talk, discussion, quarrel (implies <activity> and <sd>, *entrevista, lero-lero*)
<temp> temporal object, point in time (*amanhecer, novilúnio*, test: *até+*, cf. <dur> and <per>)
<therapy> therapy (also <domain> and <activity>, *acupuntura, balneoterapia*)

Tool prototypes:

<tool> tool, umbrella tag (*abana-moscas, lápis, computador, maceta*, "handable", cf. <mach>)
<tool-cut> cutting tool, knife (*canivete, espada*)
<tool-gun> shooting tool, gun (*carabina, metralhadora, helicanão*, in Dangram: <tool-shoot>)
<tool-mus> musical instrument (*clavicórdio, ocarina, violão*)
<tool-sail> sailing tool, sail (*vela latina, joanete, coringa*)

cp. also <mach> (machine)

<tube> tube object (*cânula, gasoduto, zarabatana*, shape-category, typically with another category, like <an> or <tool>)
<unit> unit noun (always implying <num+>, implied by <cur> and <dur>, e.g. *caloria, centímetro, lúmen*))

Weather prototypes:

<wea> weather (states), umbrella tag (*friagem, bruma*)
<wea-c> countable weather phenomenon (*nuvem, tsunami*)
<wea-rain> rain and other precipitation (*chuvisco, tromba d'água, granizo*)
<wea-wind> wind, storm (*brisa, furacão*)

SEMANTIC TAGS FOR PROPER NOUNS (HAREM tags in parenthesis)

Person categories <hum>, HAREM PESSOA:

<hum> (INDIVIDUAL) person name (cp. <H>)
<official> (CARGO) official function (~ cp. <Htitle> and <Hprof>)
<member> (MEMBRO) member

Organisation/Group categories <org>, HAREM ORGANIZACAO:

<admin> (ADMINISTRACAO, ORG.) administrative body (government, town administration etc.)
<org> (INSTITUICAO/EMPRESA) commercial or non-commercial, non-administrative non-party organisations (not place-bound, therefore not the same as <Linst>)
<inst> (EMPRESA) organized site (e.g. restaurant, cp. <Linst>)
<media> (EMPRESA) media organisation (e.g. newspaper, tv channel)
<party> (INSTITUICAO) political party
<suborg> (SUB) organized part of any of the above

currently unsupported: **<company>** (EMPRESA) company (not site-bound, unlike <inst>, now fused with. <org>)

Group categories, HAREM PESSOA:

<groupind> (GROUPOIND) people, family
<groupofficial> (GROUPOCARGO) board, government (not fully implemented)

currently unsupported: **<grouporg>** (GROUPOMEMBRO) club, e.g. football club (now fused with <org>)

Place categories <top>, HAREM LOCAL:

<top> (GEOGRAFICO) geographical location (cp. <Ltop>)
<civ> (ADMINISTRACAO, LOC.) civitas (country, town, state, cp. <Lciv>)
<address> (CORREIO) address (including numbers etc.)
<site> (ALARGADO) functional place (cp. <Lh>)
<virtual> (VIRTUAL) virtual place
<astro> (OBJECTO) astronomical place (in HAREM object, not place)

suggested: **<road>** (ALARGADO) roads, motorway (unlike <address>)

Event categories <occ>, HAREM ACONTECIMENTO:

<occ> (ORGANIZADO) organised event
<event> (EVENTO) non-organised event
<history> (EFEMERIDE) one-time [historical] occurrence

Work of art/product categories <tit>, HAREM OBRA:

<tit> (REPRODUZIDO) [title of] reproduced work, copy
<pub> (PUBLICACAO) [scientific] publication
<product> (PRODUTO) product brand
<V> (PRODUTO) vehicle brand (cp. <V>, <Vair>, <Vwater>)
<artwork> (ARTE) work of art

Abstract categories <brand>, HAREM ABSTR*ACC*AO:

<brand> (MARCA) brand
<genre> (DISCIPLINA) subject matter
<school> (ESCOLA) school of thought
<idea> (IDEA) idea, concept
<plan> (PLANO) named plan, project
<author> (OBRA) artist's name, standing for body of work
<absname> (NOME)
<disease> (ESTADO) physiological state, in particular: disease

Thing categories <common>, HAREM COISA:

<object> (OBJECT) named object
<common> (OBJECT) common noun used as name
<mat> (SUBSTANCIA) substance
<class> (CLASSE) classification category for things
<plant> (CLASSE) plant name
<currency> (MOEDA) currency name (also marked on the number)

Time categories (if used for NUM or N rather than PROP, marked only on the numeral or noun, without MWE'ing, unlike **HAREM TEMPO**):

<date> (DATA) date
<hour> (HORA) hour
<period> (PERIODO) period
<cyclic> (CICLICO) cyclic time expression

Numeric value categories (marked only on the numeral, without MWE'ing, unlike **HAREM VALOR**):

<quantity> (QUANTIDADE) simple measuring numeral
<prednum> (CLASSIFICADO) predicating numeral
<currency> (MOEDA) currency name (also marked on the unit)

Apêndice C. Conjunto de Regras de Associação

ID	Regras de Associação			
A1	hum	Hprof		
A2	HH	Hprof		
A3	Lciv	civ		
A4	Azo	A		
A5	Hfam	Hprof		
A6	Adom	A		
A7	sem-r	activity		
A8	Hfam	hum		
A9	HH	hum		
A10	Hfam	hum	Hprof	
A11	dur	per		
A12	process	ac		
A13	am	ac		
A14	Acell	meta		
A15	Acell	cc		
A16	Adom	Azo		
A17	A	meta		
A18	HH	hum	Hprof	
A19	pub	sem-r		
A20	V	Vair		
A21	act-d	activity		
A22	mat	cc		
A23	Lstar	cc		
A24	Hfam	HH		
A25	coll	cc		
A26	act-s	act		
A27	sem-c	ac		
A28	act	activity		
A29	act	ac		
A30	an	meta		
A31	cm	cc		
A32	meta	ac		
A33	meta	cc		
A34	Hfam	HH	Hprof	
A35	amount	activity		
A36	Hsick	H		
A37	therapy	activity		
A38	Ltop	top		
A39	cm-gas	cm-chem		
A40	absname	Adom		

A41	absname	A		
A42	BB	L		
A43	Lwater	object		
A44	V	object		
A45	Vair	object		
A46	org	HH		
A47	cm-chem	cm		
A48	cm-chem	cc		
A49	am	act		
A50	tool	object		
A51	mat	cm		
A52	mat	activity		
A53	event	ac		
A54	event	cc		
A55	H	HH		
A56	H	hum		
A57	coll	meta		
A58	coll	ac		
A59	object	cc		
A60	act-s	sem-c		
A61	act-s	sem-r		
A62	sem-c	act		
A63	an	cc		
A64	L	civ		
A65	Azo	meta		
A66	Azo	cc		
A67	A	cc		
A68	activity	ac		
A69	meta	HH		
A70	absname	Adom	A	
A71	V	Vair	object	
A72	am	act	ac	
A73	Hfam	HH	hum	
A74	Adom	Azo	A	
A75	Hfam	HH	hum	Hprof
A76	conv	act-s		
A77	cc-board	mat		
A78	cc-board	cc		
A79	Hattr	HH		
A80	food	fruit		
A81	unit	f-q		
A82	party	HHparty		
A83	coll-cc	cc		

A84	percep-w	ac		
A85	cord	cc		
A86	f-c	ac		
A87	Lh	org		
A88	part	an		
A89	amount	cc		
A90	Aent	A		
A91	Aent	meta		
A92	drink	cm-liq		
A93	Hbio	H		
A94	absname	Azo		
A95	cc-stone	mat		
A96	cc-stone	cm		
A97	cc-stone	Azo		
A98	cc-stone	meta		
A99	cc-stone	cc		
A100	cm-rem	cm		
A101	cm-rem	meta		
A102	cm-liq	cm		
A103	top	L		
A104	sick	meta		
A105	sick	cc		
A106	Hnat	H		
A107	Hnat	HH		
A108	Hnat	hum		
A109	Hnat	Hprof		
A110	site	org		
A111	site	HH		
A112	org	L		
A113	act-d	act-s		
A114	cm-chem	mat		
A115	tool	cc		
A116	mat	coll		
A117	event	act		
A118	event	activity		
A119	Lstar	object		
A120	Lstar	L		
A121	act-s	ac		
A122	sem-c	sem-r		
A123	sem-c	activity		
A124	an	cm		
A125	L	Lciv		
A126	Adom	meta		

A127	Adom	cc		
A128	ac	cc		
A129	absname	act-d	Adom	
A130	absname	Adom	Azo	
A131	absname	Azo	A	
A132	cc-stone	mat	cm	
A133	cc-stone	meta	cc	
A134	sick	meta	cc	
A135	Hnat	hum	Hprof	
A136	act-s	sem-c	sem-r	
A137	sem-c	act	ac	
A138	Adom	A	cc	
A139	Azo	A	meta	
A140	Azo	A	cc	
A141	Azo	meta	HH	
A142	A	meta	cc	
A143	absname	Adom	Azo	A
A144	H	HH	hum	Hprof
A145	Azo	A	meta	cc

Apêndice D. Conjunto de Regras de Classificação

C1	Hprof →hum	**C20**	Adom → Azo
C2	activity →sem-r	**C21**	Cm → cc
C3	ac →process	**C22**	Lwater → object
C4	Lciv →civ	**C23**	Adom → A
C5	H →Hsick	**C24**	Hprof ^ hum →HH
C6	act-s →act	**C25**	Coll →cc
C7	L →civ	**C26**	Site →org
C8	civ →Lciv	**C27**	an →cm
C9	per →dur	**C28**	sem-c →ac
C10	Vair →V	**C29**	occ →HH
C11	object →tool	**C30**	process →ac
C12	pub →sem-r	**C31**	Acell →cc
C13	cm-chem →cc	**C32**	H →Hbio
C14	act ^ ac →am	**C33**	Azo →Adom
C15	A →azo	**C34**	L →BB
C16	meta →A	**C35**	am →ac
C17	cc →an	**C36**	mat →cm
C18	hum →Hprof	**C37**	cm-chem →cm-gas
C19	HH →Hprof		

Apêndice E. Exemplos de anotação RST inconsistente[30]

Como mencionado anteriormente, foram encontradas algumas inconsistências de estruturação RST nos textos analisados. Essas se devem, principalmente, à atribuição de relações retóricas inadequadas aos segmentos textuais e à segmentação de EDUs. A seguir, encontram-se alguns trechos de estrutruras RST que sugerem que a anotação retórica feita por dois especialistas e, portanto, considerada de referência, seja revista para esse *corpus*.

Como se observa na Figura 22, a relação retórica utilizada para interconectar as EDUs 7-11 à EDU 6 se mostra inadequada, já que o conteúdo textual de 7-11 não traz uma explicação para o fato de 'Plutão ser o planeta mais distante até hoje fotografado', mas somente elabora a ideia expressa em 6.

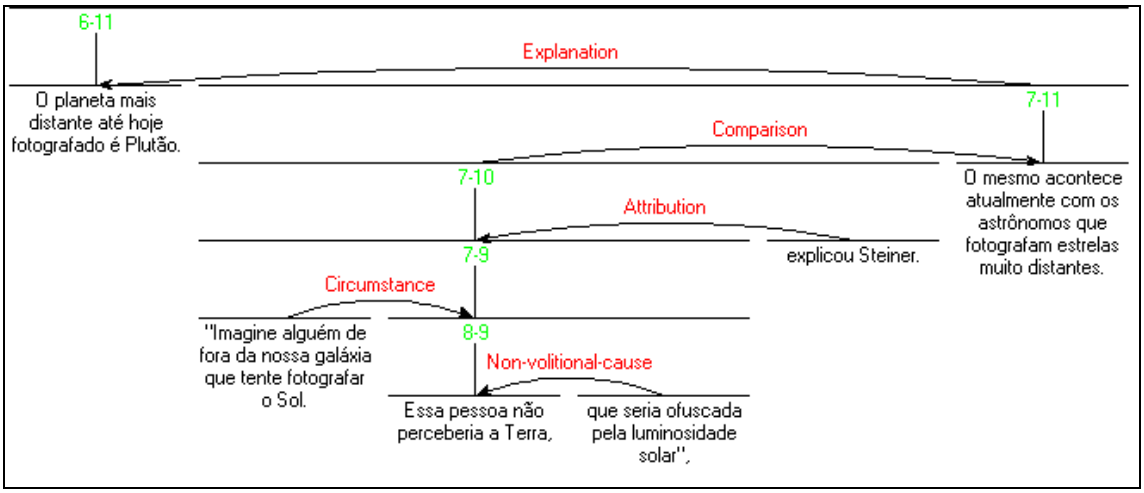


Figura 22. Trecho da estrutura RST do texto CIENCIA_2000_6380

Já na Figura 23, a seguir, o segmento expresso pela EDU 6 'no entanto há restrições' não consiste de uma interpretação do conteúdo da EDU 7 'o patenteamento só pode ser aplicado em pesquisas ligadas a doenças genéticas em que o funcionamento do gene é detalhado', mas sim de uma concessão para que o patenteamento possa ser aplicado.

[30] Neste trabalho, somente parte das estruturas RST são visualizadas, porém as estruturas completas fazem parte do pacote do *Corpus* Summ-it e encontram-se disponíveis para download no endereço: http://www.nilc.icmc.usp.br:8180/portal/index.jsp?option=downloads.jsp.

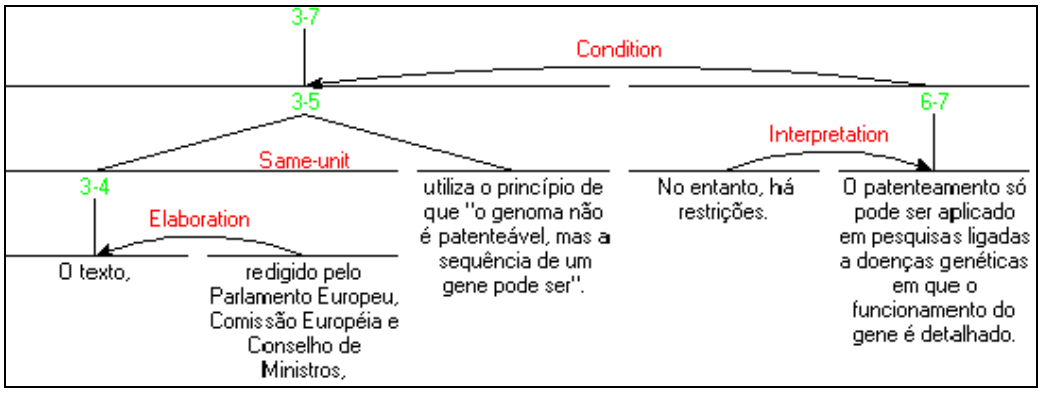


Figura 23. Trecho da estrutura RST do texto CIENCIA_2000_6381

Um texto que também apresenta inconsistência é o CIENCIA_2000_17088, com os segmentos em questão ilustrados na Figura 24 a seguir.

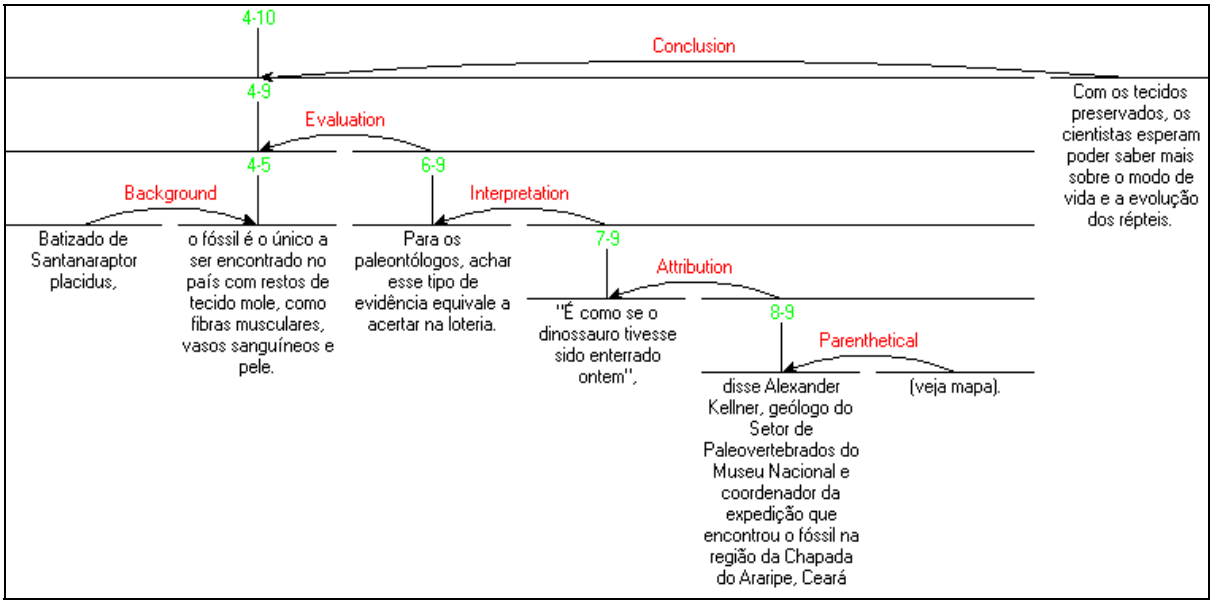


Figura 24. Trecho da estrutura RST do texto CIENCIA_2000_17088

A relação de BACKGROUND prevê que o leitor não entenderá suficientemente o N antes de ler o S, portanto observa-se que a informação constante da EDU 4 'Batizado de Santaraptor placidus' não deveria receber essa relação retórica, mas sim o de uma ELABORATION, já que simplesmente dá nome ao fóssil do dinossauro encontrado. Além disso, outro problema de mesma natureza foi encontrado neste mesmo trecho: o conteúdo da EDU 7 'É como se o dinossauro tivesse sido enterrado ontem" não indica uma interpretação da EDU 6, mas uma comparação com 'ganhar na loteria', pois o fóssil ainda preserva restos de

tecido mole e até pele. Tal comparação é indicada inclusive pelo marcador discursivo ‘como’na superfície textual.

Em outros casos, o problema se encontra na segmentação das unidades textuais, como no caso da EDU 14 da Figura 25, expressa pelo segmento ‘alegando’, o qual não expressa uma ideia completa para que seja interconectada a outra EDU. A sugestão, neste caso, é que o conteúdo das EDUs 13 e 14 formassem uma única EDU, que aí, sim, poderia estar ligada à subárvore que a segue.

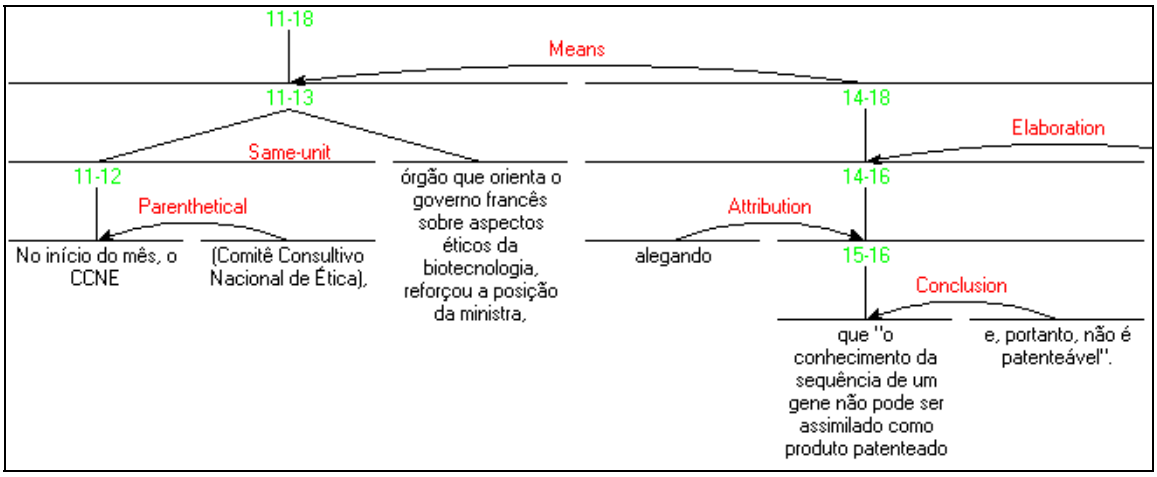


Figura 25. Trecho da estrutura RST do texto CIENCIA_2000_6381

No caso da Figura 26, o questionamento se deve à quebra do segmento ‘produção de mudas em laboratório feita para evitar doenças e selecionar vegetais saudáveis’, o qual deveria ser expresso pela relação PARENTHETICAL, porém foi desmembrado em vários outros segmentos e novas relações foram atribuídas, consequentemente. Para o fim de sumarização, essa quebra do segmento poderia comprometer a mensagem final veiculada.

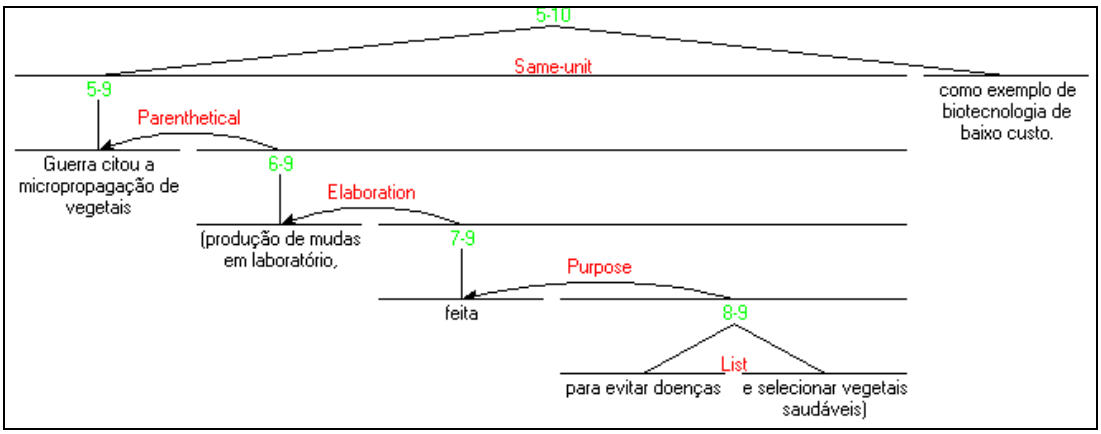


Figura 26. Trecho da estrutura RST do texto CIENCIA_2000_6389

A partir dos exemplos apresentados nesta seção, justifica-se a necessidade de revisão das estruturas RST consideradas de referência, para que elas sejam adequadas tanto em sua segmentação textual, quanto na utilização de relações retóricas.