

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



Anotação Parcial de Estruturas Retóricas (RST) do Corpus Summ-it

Thiago Ianez Carbonel
Juliana Thiesen Fuchs
Lucia Helena Machado Rino

NILC-TR-07-04

Maio, 2007

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

Resumo

Apresentamos, neste relatório técnico, os dados obtidos na tarefa de anotação retórica de parte do Corpus Summ-it (corpus de textos jornalísticos de divulgação científica). A teoria RST – *Rhetorical Structure Theory* – foi usada para produzir as estruturas retóricas, ou estruturas RST. Introduzimos, previamente, o arcabouço teórico que dá suporte ao trabalho – noções de textualidade e da própria RST – para, em seguida, detalharmos a metodologia de anotação, especificando as orientações de segmentação dos textos e de escolha das relações retóricas. Constituem resultados deste trabalho, ao final, um conjunto de considerações relativas a padrões de escolhas retóricas e seus respectivos marcadores indicativos que, pelo reiterado uso nos textos analisados, foram apontados como indicativos de uma possível tipologia para o gênero jornalístico sob enfoque.

Este trabalho conta com o apoio financeiro do CNPq.



Índice geral

1.	Introdução	5
2.	Textualidade	7
2.1.	Noção de Texto.....	7
2.2.	Coesão textual.....	11
2.3.	Coerência textual	17
3.	Teoria de Estruturação Retórica (Rhetorical Structure Theory – RST)	19
4.	Metodologia de anotação RST do Corpus Summ-it.....	24
4.1.	O protocolo de anotação RST.....	25
4.2.	Segmentação proposicional	30
4.2.1.	Segmentação proposicional do Corpus Summ-it.....	31
4.2.2.	Problemas de Segmentação do Corpus Summ-it.....	31
4.3.	Estruturação retórica.....	33
4.3.1.	Pressupostos teóricos e práticos	33
4.3.2.	Problemas de Estruturação do Corpus Summ-it.....	35
5.	Resultados da análise do corpus parcial	39
6.	Considerações finais	46
	Referências bibliográficas	47
	Anexo A – Textos originais do Corpus Summ-it	49

Índice de figuras

Figura 1. Sentença (A) e sua estrutura RST.....	21
Figura 2. Sentença (B) e sua estrutura RST.....	21
Figura 3. Definição da relação PURPOSE.....	21
Figura 4. Definição da relação SEQUENCE.....	21
Figura 5. Segmento ilustrativo do texto CIENCIA_2003_24219.....	22
Figura 6. Estrutura RST parcial do texto (C).....	22
Figura 7. Ilustração da relação EVIDENCE.....	27
Figura 8. Árvore RST do texto CIENCIA_2000_17109	35
Figura 9. ELABORATION entre EDUs do texto CIENCIA_2000_17108.....	36
Figura 10. NON-VOL. RESULT entre EDUs do texto CIENCIA_2000_17108	37
Figura 11. Alteração da estrutura ilustrada na Figura 10	37
Figura 12. EXPLANATION entre 1-2 e 3-4 do texto CIENCIA_2000_17101.....	38
Figura 13. CONTRIBUTION entre 1-2 e EDU3 do texto CIENCIA_2005_28756	39
Figura 14: Análise do texto CIENCIA_2000_17109.....	43
Figura 15. Análise de parte do texto CIENCIA_2000_17082.....	44
Figura 16. Finalização estrutural do texto CIENCIA_2005_28747.....	466

Índice de tabelas

Tabela 1. Conjunto original de relações RST.....	23
Tabela 2. Incidência das relações RST nos 12 textos do corpus Summ-it.....	40
Tabela 3. Incidência de relações no corpus	41

1. Introdução

Com a crescente expansão das vias digitais de difusão da informação (em especial a Internet), tem sido cada vez maior a demanda por formas alternativas de condensação dos documentos distribuídos eletronicamente. A idéia de produzir um texto contendo apenas as informações centrais a partir de um texto mais elaborado se coaduna perfeitamente com esta tendência global.

Sumarizar é uma atividade essencialmente humana. Nos atos de comunicação mais triviais os indivíduos constroem representações condensadas da informação que transmitem aos seus interlocutores, seja ao comentar sobre uma palestra, seja contando um episódio de uma novela. O processo de sumarização humana, apesar de ser uma atividade cognitiva, pode ser descrito em termos de etapas mimetizáveis por um sistema computacional. O leitor humano, ao defrontar-se com um texto, realiza vários níveis de leitura que lhe permitem identificar a idéia central e elencar quais são as informações nucleares com relação à mesma. A premissa para a automação do processo é a possibilidade de que um sistema de processamento de língua natural (doravante PLN) faça o mesmo.

A Sumarização Automática de textos (doravante SA) consiste na atividade de geração automática de uma versão condensada de um dado texto-fonte (Mani, 2001). Assim sendo, um bom sumário deve conter a idéia principal de seu texto-fonte, articulando em torno dela suas informações centrais. Este processo, na esteira das demandas atuais, tem amplas aplicações, que vão desde a escrita científica (geração automática de *abstracts*, por exemplo) até a incorporação a mecanismos de busca na Internet, nos quais a SA pode auxiliar na identificação de documentos relevantes.

Os sumários gerados atualmente, porém, ainda estão distantes dos resultados obtidos por humanos, ou seja, do ideal. Muitos são os problemas que podem ser elencados: presença de informações redundantes ou irrelevantes, agramaticalidade e déficits de textualidade em geral, ou seja, problemas de coerência e coesão, em especial a perda da referência por conta de quebras de elos referências geral, ou seja, problemas de coerência e coesão. Um problema comum, em especial, é a perda da referência por conta de quebras de elos referenciais presentes no texto-fonte.

Segundo Sparck Jones (1999), os modelos de SA podem ser divididos em duas categorias básicas: os pobres em conhecimentos lingüísticos (modelos estatísticos, ou seja, abordagem superficial) e os que utilizam significativamente estes conhecimentos (abordagem profunda). Atualmente, apesar da disponibilidade e utilização de modelos estatísticos, são os modelos de abordagem profunda que permitem a manipulação de dados textuais com vistas à solução de problemas como a manutenção de elos co-referenciais e as questões de textualidade dos sumários gerados.

O trabalho de Seno (2005) sinaliza que, a partir da utilização de teorias lingüísticas, é possível proceder à automação do processo de sumarização, levando em consideração características lingüísticas do texto-fonte (sua estrutura retórica) e utilizando heurísticas de poda baseadas em dois modelos distintos: aqueles que calculam a saliência da informação (Marcu, 1997) e aqueles possibilitam a determinação do que se denomina “domínio de acessibilidade referencial” (Cristea et al., 1998).

Deste modo, o processamento automático das cadeias de co-referência (com vistas especialmente à SA), requer a utilização de modelos lingüísticos, e é nesse contexto que se insere o Projeto ProCaCoSA (CNPq, no. 503766/2005-4). A tarefa que descrevemos neste relatório corresponde à preparação de um corpus coletado para a pesquisa. Nesta etapa, dois anotadores especialistas (lingüistas) procederam à análise das estruturas retóricas dos textos, utilizando a Teoria de Estruturação Retórica (Mann & Thompson, 1987), mais conhecida por sua sigla na língua inglesa, RST. O foco de estudo apresentado é a discussão das peculiaridades da anotação¹ manual, em todos os seus níveis (segmentação dos textos, escolhas de relações etc.), a fim de possibilitar ao grupo uma base para discussões e sugestões.

Na seção 2, apresentamos noções elementares de Lingüística Textual, relevantes aos propósitos do projeto no qual esta anotação se insere e, na seção 3, introduzimos a teoria RST. Na seção 4, detalhamos nossa metodologia de anotação do corpus e, na seção 5, enfim, apresentamos alguns resultados preliminares que podem ser depreendidos deste trabalho.

¹ Trata-se do termo usual para o processo de análise com vistas à produção de um outro texto, o anotado com informações retóricas.

2. Textualidade

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

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

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

2.1. Noção de Texto

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

Durante muito tempo, não houve uma preocupação pontual acerca da elaboração de um ramo da ciência que tivesse o texto como objeto de análise – o texto era um meio e não um fim. Foi só a partir de meados do século XX que se iniciou um movimento de elaboração de gramáticas textuais. Nessa fase, o texto era

considerado apenas uma estrutura lingüística organizada – falando-se então em *texto* (constituintes lingüísticos *coerentemente* organizados) e em *não-texto* (constituintes lingüísticos organizados *sem coerência*). Segundo Koch (2004), nesta primeira fase, os conceitos de texto variaram desde "unidade lingüística (do sistema) superior à frase" até "complexo de proposições semânticas".

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

Neste trabalho, é importante salientar, as denominações *texto* e *discurso* são tomadas como sinônimos ainda que, no contexto atual dos estudos lingüísticos, tal confusão já esteja suficientemente elucidada. Nossas razões são de cunho essencialmente pragmático e levam em conta a cristalização do termo *discurso* nos estudos lingüístico-computacionais. Assim, cumpre estabelecer os parâmetros conceituais adotados neste trabalho, definindo não apenas *texto* (tendo *discurso* como sinônimo), mas também *gênero textual* e *tipo de texto*, que são conceitos utilizados de modos diferentes na Lingüística e na Lingüística Computacional.

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

Gêneros diferentes de texto, com prerrogativas distintas, foram determinados para se separar o texto (produto textual) do não-texto (produto prejudicado pela má elaboração). Essa assunção passou a ser extremamente importante para várias pesquisas em Lingüística Computacional.

Segundo Bonini (2001), entre as abordagens mais conceituadas sobre gênero textual, na atualidade, estão as de Van Dijk (1979), Swales (1992), Biber (1988) e Bronckart (1999). À parte as peculiaridades, as três primeiras, embora relacionando ao contexto social de origem e associando propósitos comunicativos, concebem o gênero textual como uma estrutura composta de partes características, agrupadas sob determinada sintaxe que reproduz uma ordem canônica. Partem, no geral, de uma descrição do texto conforme é caracterizado e utilizado em determinado ambiente social, mas de forma generalizante e com a atenção voltada para os aspectos formais. Nesse sentido, De Beaugrande e Dressler (1981) afirmam que existem correspondências regulares entre a estrutura de um texto e a estrutura do mundo que o texto evoca. Esse conceito se aplica à estrutura geral que caracteriza a exteriorização do pensamento técnico-científico em sua forma mais abrangente.

Neste trabalho, adotamos o conceito de *gênero* proposto por Swales (1992, p. 58), que encerra as considerações feitas acima²:

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

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

² Tradução e ênfases nossas.

Tomemos o caso do analisador discursivo automático – DiZer (Pardo et al., 2004; Pardo, 2005) – que trata textos científicos, produzindo sua estrutura retórica. Em seu processamento são consideradas palavras e expressões indicativas como ponteiros (indicadores) para a interpretação automática. Elas são elementos constitutivos do texto que podem ser facilmente identificáveis automaticamente, como verbos que indicam relações ou frases cristalizadas (por exemplo, “o objetivo deste trabalho é”). Todavia, para gêneros diferentes de texto esses elementos podem variar ou, mesmo que permaneçam comuns entre gêneros, podem apontar relações retóricas diferentes, prejudicando o desempenho do sistema.

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

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

Em vista dessas considerações, optou-se, neste trabalho, pelo conceito de texto mais amplo delineado por Koch (2004, p. 18), cujo trabalho é de suma relevância nos estudos da Lingüística Textual³:

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

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

2.2. Coesão textual

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

A esta organização linear dos constituintes lingüísticos do texto dá-se o nome de *coesão*, que pode ser definida, sem muitas variações entre os autores, como sendo o uso de meios lingüísticos para facilitar a coerência (Halliday & Hasan, 1976; De Beaugrande e Dressler, 1981). Um elemento de coesão indica como duas porções de

³ Ênfases nossas.

texto ligam-se conceitualmente uma à outra – a essa ligação dá-se o nome de *elo coesivo*.

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

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

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

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

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

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

"O presidente da Venezuela, Hugo Chavez, pediu às famílias venezuelanas Em seu programa semanal de rádio transmitido no domingo, ele disse que o *Halloween* é um jogo do terror, segundo a agência de notícias *Associated Press*."

O principal problema desse sumário é decorrente da dificuldade de processamento automático e é o seguinte: a primeira sentença do sumário está incompleta em razão da falta do complemento verbal direto [o que o agente da ação pediu]. Este não foi produzido por duas razões: o sumarizador automático o considerou uma unidade independente da anterior e pouco relevante para inserção no sumário. Como consequência de sua ausência, é praticamente impossível ao leitor recuperar o que o escritor queria dizer realmente, ou o que foi que o presidente Hugo Chavez pediu às famílias. Essa má estruturação implica, assim, a deficiência do sumário, no que se refere à textualidade. No entanto, como o restante do texto transmite uma idéia completa, seria possível que traçássemos algumas hipóteses para a completeza da primeira sentença (e decorrente correção do déficit), na tentativa de identificá-la como parte da mensagem pretendida pelo escritor. Uma delas seria a de que um leitor humano poderia depreender com certa segurança que o pedido seria para que as famílias venezuelanas não participassem ou compactuassem com o *Halloween* (H1). Outra, a de que o pedido seria para que as famílias dessem crédito à notícia veiculada pela *Associated Press* (H2).

Configurando-se H1 ou H2, não seria o caso, necessariamente, de considerar que a textualidade do texto é tão prejudicial que não seria possível recuperar alguma mensagem, já que o leitor poderia assumir uma das duas hipóteses. É por isso que configuramos esse problema como um déficit (opostamente a impedimento de compreensão). Mais particularmente, esse déficit se deve à má estruturação textual introduzida pela amarração "frouxa" entre os constituintes lingüísticos do texto, introduzida pela omissão de um complemento importante na primeira sentença.

Esse exemplo sugere questões igualmente importantes para a produção e interpretação textual: problemas expressos em um texto podem se manifestar em vários níveis – desde a ausência de informação, a qual pode acarretar sua má estruturação, requerendo do leitor sua capacidade de fazer suposições (inferências) para deduzir o sentido, até a completa impossibilidade de recuperação do sentido

pretendido pelo autor. Mais importante, a textualidade não se confunde com a capacidade de, a partir da trama textual, recuperar-se o sentido do texto. Seria perfeitamente possível termos um texto mal escrito no qual o leitor, ainda assim, pudesse recuperar a informação que o autor inicialmente pretendeu transmitir. Porém, se sua recuperação demandar um esforço expressivo, configurado no exemplo acima pela necessidade de se estabelecer uma das duas hipóteses, a compreensão e, assim, a aceitação do texto e a comunicação de sua mensagem ficam totalmente comprometidas.

Queremos mostrar, com esse exemplo, que problemas gerados automaticamente não necessariamente levam a problemas de entendimento do produto final. No entanto, estabelecer um modelo computacional que consiga assegurar o tratamento dessas nuances para os problemas de textualidade não é trivial. Veremos que, no nosso caso, elas são tratadas em um contexto muito particular que pode ser entendido, sobretudo, pelos estudos de Halliday e Hasan (1976). Segundo eles, a coesão ocorre quando dois elementos no discurso são interdependentes, ou seja, a interpretação de um deles depende da interpretação do outro. Um pressupõe o outro, no sentido de que não pode ser efetivamente decodificado a não ser por recurso ao outro. Trata-se de um conceito semântico de coesão, que explora as relações de sentido internas ao texto e responsáveis pela constituição do mesmo enquanto um texto.

Assim, retomando o exemplo anterior, podemos identificar relações entre os constituintes lingüísticos do texto (elementos do discurso) que nos permitiriam, mesmo a partir do problema apontado (falta do argumento do verbo ‘pedir’), inferir o sentido do texto. Semanticamente, o texto destacado poderia ser expresso segundo os seguintes constituintes:

{	A (agente) ‘pedir’ (verbo – ação) B (argumento indireto – destinatário do pedido) C (argumento direto – o que é pedido) D (justificativa do pedido)	}
---	-----------------------------------------------------------------------------------------------------------------------------------------------------------------	---

A construção do tecido textual, segundo os autores, ocorre na medida em que vão se estabelecendo relações de sentido (semânticas) entre as sentenças encadeadas. A união dessas sentenças forma o que eles denominam de elo coesivo. Desse modo, consideremos o seguinte exemplo:

[“O Supremo Tribunal Federal manifestou-se favoravelmente à pesquisa com células-tronco no Brasil.”]₁ [**Com isso**, a perspectiva de investimentos em pesquisa pública e privada deve sofrer alterações drásticas”]₂

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

[“O Supremo Tribunal Federal manifestou-se favoravelmente à pesquisa com células-tronco no Brasil.”]₁ [**No entanto**, a perspectiva de investimentos em pesquisa pública e privada deve sofrer alterações drásticas”]₂

A informação em (1) mantém-se inalterada, mas (2) sofreu uma modificação: o marcador, agora, não mais introduz uma relação de decorrência lógica, mas sim de adversidade ou contraste: (2) é uma situação em sentido contrário ao que se compreende em (1). Como consequência, a interpretação que se dá a [alterações drásticas], agora, deve acompanhar o sentido relacional introduzido pelo marcador e, assim, levar a um efeito negativo das alterações, ou seja, a uma diminuição de investimentos.

Para Halliday e Hasan, portanto, a coesão e a coerência estão intimamente ligadas, uma vez que a segunda depende da adequação da primeira. Contrário a esse conceito de coesão textual, Marcuschi (1983) busca distinguir a coesão da coerência, afirmando serem conceitos separados. Segundo ele, é possível haver textos cuja organização linear se dê apenas no nível do sentido, e não pela intermediação de constituintes lingüísticos. Um exemplo seria o seguinte texto:

(a)

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

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

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

(b)

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

O segmento acima possui uma estruturação coesiva evidenciada pela presença dos marcadores – então, por causa disso, e – mas, por faltar-lhe qualquer sentido lógico, ou seja, por ser incoerente, não se pode, em princípio, denominá-lo de texto⁴.

O fato de ser possível a existência de textos desprovidos de elementos coesivos não é razão para sua não-valorização. Se, por um lado, existe a referida possibilidade – e tais textos pertencem a um domínio bastante restrito (literatura, texto telegráfico etc.) – a coesão marcada por elementos explícitos confere ao texto maior legibilidade (Koch, 2004), sendo altamente desejável – quando não indispensável – na maior parte dos gêneros textuais (científico, didático, científico, jornalístico etc.). Isso porque a coesão é uma propriedade que permite explicitar as relações existentes entre os constituintes do texto, possibilitando, assim, maior grau de compreensão ao reduzir as possibilidades de leituras equivocadas.

⁴ Ressalvada, aqui, a hipótese de se tratar de texto poético. A linguagem poética possui uma liberdade criativa cuja complexidade foge ao círculo no qual este trabalho está delimitado.

Os recursos lingüísticos disponíveis para a estruturação reticular do texto são variados. Koch sistematiza o estudo destes mecanismos dividindo-os em duas modalidades: os de coesão seqüencial e os de coesão referencial, sendo esta última o foco deste estudo.

Koch define a *coesão referencial* como aquela em que um componente da superfície do texto faz remissão a outros elementos nela presentes ou inferíveis a partir do universo textual. A definição da autora engloba duas entidades referenciais distintas que devem ser identificadas e mais especificamente definidas a fim de que a distinção reste clara. Quando falamos em remissão de um componente do texto a outro nele presente, temos uma **forma referencial**, que também é chamada de **forma remissiva**. Isso porque, no corpo do texto, é possível apontar uma entidade concreta como o termo antecedente à expressão referencial – a este termo antecedente dá-se o nome de **elemento de referência**, ou como Koch o denomina, um **referente textual**. O segundo caso descrito, quando um componente textual é compreensível somente por processos inferenciais, o elemento de referência (antecedente) não se encontra no texto, mas sim no contexto. Em que ele se insere. Nos estudos situados no campo da Lingüística Computacional, o termo comumente utilizado para se referir a esse fenômeno é **cadeia de co-referência**.

O fenômeno do uso de referência na construção de um texto é amplo e precisa de um recorte preciso que permita um estudo pontual do assunto. As referências que constroem os sentidos de um texto podem ser, portanto, tanto internas (endofóricas) quanto externas (exofóricas). As referências externas ao texto são aquelas que servem de sustentáculo para a compreensão do texto e, em geral, correspondem ao contexto de leitura. No trabalho apresentado neste relatório, detemo-nos apenas nas referências endofóricas do tipo anafóricas, ou seja, aquelas posteriores ao termo referente.

2.3. Coerência textual

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

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

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

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

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

Vimos nesta seção uma noção funcional para as aplicações de textualidade e seus mecanismos neste trabalho. No que segue, veremos o modelo de representação do discurso adotado, o modelo de estruturação retórica. Antes, porém, ressaltamos que é importante considerar que é senso comum na Lingüística Textual, hoje, que a textualidade é construída por um conjunto de fatores diversos, intra e extralingüísticos, o qual ainda escapa a uma enumeração precisa. De Beaugrande e Dressler (1981) propõem sete fatores fundamentais: coesão, coerência, informatividade, situacionalidade, intertextualidade, intencionalidade e aceitabilidade. Desses, apenas a coesão e a coerência estão adstritas aos limites intratextuais, razão pela qual são de maior interesse para sistemas que operam dados computáveis, como é o caso da Sumarização Automática. No entanto, é importante salientar que as iniciativas de Sumarização Automática, como produção textual, buscam contemplar a informatividade, a intertextualidade (situacionalidade), a intencionalidade e a aceitabilidade dos sumários gerados automaticamente. Por se tratarem, porém, de abordagens computacionais, os modelos que buscam assegurar

essas propriedades são diversificados, podendo mesmo ser inexistentes. Procurar assegurar durante a produção de um sumário que ele seja aceitável, por exemplo, exige que se tenha um modelo artificial da aceitabilidade, o que implica a necessidade de um *modelo do usuário* (ou da audiência suposta). Somente dessa forma é possível definir a aceitabilidade como um parâmetro dependente do modelo que se supõe fundamentar um sistema automático. Além disso, uma vez adotado um modelo desse tipo, ele tem que ser mantido em todas as etapas que dele dependem, para que haja consistência teórica e prática, levando a resultados pertinentes. Por exemplo, ele tem que ser usado em todas as etapas de avaliação de sumários gerados automaticamente pelo sistema que tem esse modelo como base. Essas questões de longe são triviais e, assim, na maioria das vezes, essas propriedades não se materializam como parâmetros de um sistema, levando-o a uma caracterização genérica e, portanto, menos dependente de modelos particulares ou sofisticados.

3. Teoria de Estruturação Retórica (Rhetorical Structure Theory – RST)

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

A RST fundamenta-se no princípio de que um texto tem uma estrutura retórica subjacente à estrutura superficial. Através dessa estrutura retórica, é possível recuperar o objetivo comunicativo que o escritor⁵ do texto pretendeu atingir ao escrevê-lo. Embora esses pressupostos da RST remetam ao aspecto discursivo, seu uso se limita à Lingüística Textual, ou seja, discurso somente enquanto texto. Este pode ser segmentado em unidades mínimas de significado (ou conteúdo) denominadas EDUs – *Elementary Discourse Units* (Marcu, 1997) – que, necessariamente, mantêm relação entre si na construção textual. Suas relações são

⁵ Usaremos, aqui, os termos escritor-leitor em contrapartida ao binômio produtor-receptor, já que nosso foco são as manifestações verbais escritas, unicamente, as quais envolvem o escritor como produtor e o leitor como receptor.

previamente definidas pelo modelo, cujo conjunto, apesar de não definitivo, pretende ser suficientemente amplo para cobrir os casos retóricos considerados.

As relações RST são divididas em duas classes: na concepção dos autores, chamadas de mono e multi-nucleares; na concepção de Marcu (1997), de hipotáticas e paratáticas (Marcu, 1997). As relações hipotáticas inter-relacionam pares de EDUs que apresentam diferentes graus de importância, sendo uma nuclear e a outra satélite, daí o serem mononucleares. As relações paratáticas inter-relacionam EDUs que apresentam o mesmo grau de importância, o que lhes confere a denominação de multinucleares. Claramente, a diferença entre essas denominações não é casual: Mann e Thompson contemplam o nível de organização do discurso e, assim, o aspecto principal em questão é o da forma como duas EDUs se inter-relacionam estruturalmente. Já Marcu contempla o nível de importância ou saliência das mesmas, o que confere à sua denominação um caráter funcional distinto e até mesmo complementar ao anterior: EDUs nucleares serão mais salientes do que EDUs satélite, enquanto várias EDUs nucleares e inter-relacionadas serão igualmente salientes.

Como forma de ilustrar os dois tipos de relação RST, vejamos os exemplos apresentados nas Figuras 1 e 2, para os textos mono-sentenciais (A) e (B) respectivamente. Os números entre colchetes indicam a delimitação de cada EDU e os ramos em negrito nas estruturas RST, a informação nuclear (em oposição à satélite). Podemos observar em (A) que a EDU 2 introduz o propósito (*purpose*) da EDU 1, sendo esta o núcleo e a outra, o satélite da relação *purpose*. Em (B) temos uma relação de seqüência (*sequence*) entre os segmentos que não é hierárquica e, portanto, é multinuclear. As definições dessas relações na Teoria RST são apresentadas nas Figuras 3 e 4, respectivamente⁶. Optou-se, neste trabalho, por não traduzir os nomes das relações, haja vista a prática comum em outros trabalhos na área de Linguística Computacional.

⁶ Todas essas definições foram traduzidas das originais, presentes em várias obras, por exemplo, (Mann & Thompson, 1987) e (Mann et al., 1992) ou no próprio sítio da RST (<http://www.sfu.ca/rst/01intro/definitions.html>).

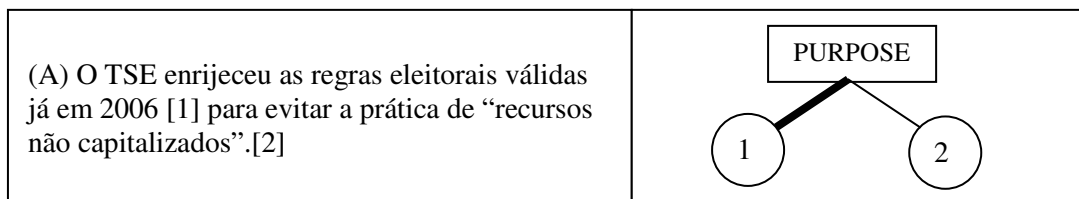


Figura 1. Sentença (A) e sua estrutura RST

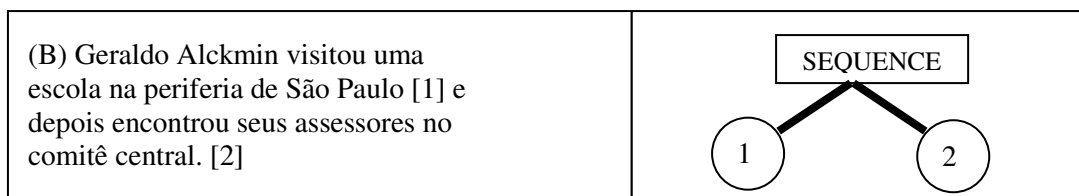


Figura 2. Sentença (B) e sua estrutura RST

<p>Nome da relação: PURPOSE</p> <p>Condições sobre núcleo (N): N apresenta uma ação</p> <p>Condições sobre satélite (S): S apresenta uma situação não realizada</p> <p>Condições sobre combinação núcleo-satélite (N+S): S deve ser realizada pela ação em N</p> <p>Efeito: o leitor reconhece que a ação em N é iniciada para realizar S</p>

Figura 3. Definição da relação PURPOSE

<p>Nome da relação: SEQUENCE</p> <p>Condições sobre os Ns: as situações apresentadas nos Ns são realizadas em sequência</p> <p>Efeito: o leitor reconhece a sucessão dos eventos apresentados no N</p>

Figura 4. Definição da relação SEQUENCE

Como mostram os exemplos acima, a estruturação RST resulta em uma árvore e pode usar quaisquer relações do conjunto de relações RST, as quais envolvem somente EDUs, isto é, unidades elementares do texto em foco. Embora os exemplos ilustrados nas figuras 1 e 2 indiquem árvores simples, com uma única relação RST, no caso geral uma árvore RST é construída composicionalmente, ou seja, relações se estabelecem também entre sub-árvores RST, como mostra a estrutura RST para o texto (C) exibido na Figura 5. Esse texto, que compõe nosso corpus de trabalho (descrito na Seção 4) e foi extraído da Folha On-line (07/11/2006), foi analisado por um especialista em RST e sua estrutura, ilustrada na Figura 6, reflete a estrutura arbórea construída manualmente com o auxílio de uma ferramenta automática, a RSTTool (O'Donnel, 2000).

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

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

Figura 5. Segmento ilustrativo do texto CIENCIA_2003_24219

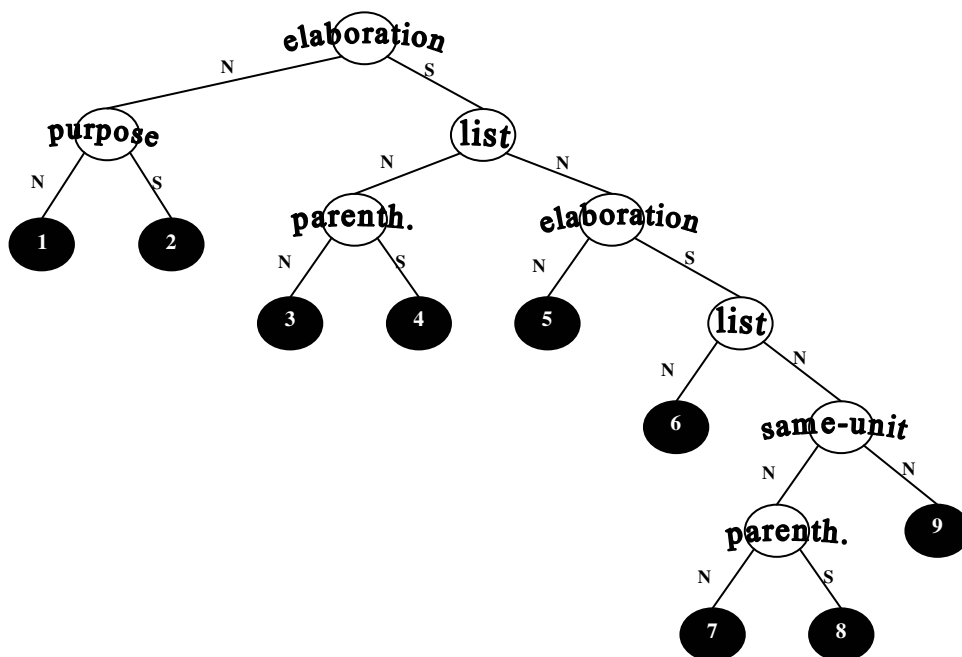


Figura 6. Estrutura RST parcial do texto (C)

Inicialmente, Mann e Thompson estabeleceram um conjunto de apenas 24 relações, as quais são exibidas na Tabela 1, juntamente com seu tipo de nuclearidade. Outros autores, como Marcu (1997), propõem conjuntos bastante diversos desse, remontando a mais de cem relações, o que torna um processo de análise bastante mais complexo. Na Figura 6, por exemplo, utilizamos a relação *parenthetical*, ausente no conjunto original de Mann e Thompson, mas presente em outros trabalhos (O'Donnell, 1997; Marcu, 1997; Pardo, 2005). Como se pode notar comparando-se essa estrutura RST com o texto na Figura 5, algumas dessas relações têm significado evidente, já que são familiares a um falante competente, como a de propósito (*purpose*) ou elaboração (*elaboration*). O mesmo se aplica a outras relações RST não ilustradas, como evidência (*evidence*) ou contraste (*contrast*). Entretanto, outras são mais obscuras, como *enablement* ou *background*.

Para compreender adequadamente os parâmetros que discriminam uma relação e recuperar seu contexto de uso, o leitor/analista deve recorrer à obra de referência.

Tabela 1. Conjunto original de relações RST

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

Para construir árvores RST como as ilustradas, o analista deve, primeiramente, reconhecer os tópicos principais de um texto e, assim, sua idéia principal, a fim de traçar o relacionamento retórico mais indicado entre os elementos macro-estruturais. Entretanto, a construção da estrutura se dá primeiramente pela segmentação do texto em EDUs e pelo seu relacionamento, ou seja, pela construção de sub-árvores simples. Desse modo, tanto a recuperação da macro-estrutura quanto a recuperação da micro-estrutura são relevantes. Ao mesmo tempo, dados os pressupostos da Teoria RST, a tarefa de análise visa recuperar as intenções do escritor. Assim, o analista deve ser, ao mesmo tempo, um leitor competente e um especialista na representação do conhecimento, para elaborar sua tarefa de modelagem artificial a fim de produzir uma estrutura RST segundo os pressupostos da Teoria.

Via de regra, o conhecimento requerido para a estruturação RST envolverá padrões de análise, conhecimento morfosintático, reconhecimento de marcadores textuais e sua correspondência com as relações retóricas. Entretanto, a descoberta de padrões de análise é altamente dependente da apreensão da mensagem, ou idéia principal, do texto, assim como de sua organização macro-estrutural. Assim, é importante também que o analista tenha conhecimento do domínio (assunto) específico sobre o qual versa o texto, sobretudo se estiver tratando domínios muito complexos ou específicos que dificultam a tarefa de recuperação do inter-relacionamento

proposicional adequado. Ao mesmo tempo, ser um usuário da teoria e um leitor competente tornam-se essenciais para o analista.

Em geral, os padrões indicativos das relações RST são dependentes de gênero e domínio textual: para sua determinação é preciso reconhecer a ordem das proposições (sejam elas EDUs ou segmentos textuais mais complexos), a qual será determinante da nuclearidade, isto é, do reconhecimento das unidades que serão núcleos e satélites, assim como de seu relacionamento funcional (determinação da relação retórica, propriamente dita).

A anotação RST consiste, portanto, na recuperação do mapeamento de uma situação comunicativa apresentada em língua natural, previamente elaborado pelo escritor. Nesse sentido, essa situação espelha o modo e as razões de haver usos particulares da língua natural. Assim, tem-se por hipóteses que: a) a língua natural e a situação discursiva levam ao *efeito do discurso sobre o leitor* – aqui é possível descobrir, por exemplo, porque os usos particulares da língua natural podem ter sucesso ou falhar ante o leitor; b) o escritor deseja provocar, com seu texto, efeitos particulares no leitor; c) o texto é fundamentado, portanto, nas intenções do escritor, conforme evidenciado pelas figuras 3 e 4.

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

4. Metodologia de anotação RST do Corpus Summ-it

O corpus anotado pelo projeto ProCaCoSA, denominado Summ-it, está diretamente ligado ao projeto PLN-BR (Recursos e Ferramentas para a Recuperação de

Informação em Bases Textuais em Português do Brasil)⁷. Ele constitui-se de um sub-corpus do corpus PLN-BR Gold⁸ e é formado por 50 textos jornalísticos retirados do caderno de Ciências da Folha de São Paulo, escritos em português do Brasil e voltados ao contexto midiático. Esse corpus foi escolhido por questões operacionais de trabalho: os textos foram, paralelamente, anotados com cadeias de co-referência. Uma vez que o interesse direto do Projeto ProCaCoSA é exatamente o processamento das referidas cadeias, era de interesse do projeto que os mesmos textos fossem também anotados com suas estruturas retóricas, processo detalhado nesta seção somente quando aplicado a uma amostra de 12 textos do corpus Summit, os quais são apresentados no Anexo A.

4.1. O protocolo de anotação RST

Como se pode depreender da Seção 3, a anotação de textos trata de uma tarefa que, apesar de orientada por um conjunto pré-estabelecido de passos, insere ainda muito da subjetividade do anotador, razão pela qual o relatório detalhado das motivações das nossas escolhas, como anotadores, torna-se um repositório importante de informações a serem utilizadas em trabalhos futuros de anotação. Mesmo revisões futuras das anotações atuais do corpus podem se beneficiar deste relatório. Esse detalhamento é de tal relevância que motivou a construção de ferramentas de suporte a esse tipo de análise, como a RhetDB ou a RST Toolkit, ambas de autoria de Thiago Pardo⁹, destinadas a lidar com estruturas RST previamente construídas com a ferramenta RSTTool, já citada.

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

⁷ Proc. CNPq Nro. PDITI 550388/2005-2, que envolve sete universidades brasileiras, cada qual com seu próprio subprojeto de PLN, sendo todos eles inter-relacionados.

⁸ Disponível em <http://nilc.icmc.usp.br:8180/portal/>.

⁹ Disponíveis para download em <http://www.icmc.usp.br/~taspardo/>.

Uma vez segmentado o texto, dois anotadores humanos, especialistas (lingüistas familiarizados com a RST), estruturaram macro e micro-estruturalmente os textos, valendo-se do auxílio da ferramenta de suporte RSTTool. Note-se que essa ferramenta não automatiza a análise em nenhum aspecto, servindo apenas como um ambiente amigável para a anotação que fornece recursos gráficos úteis para a visualização da estrutura arbórea decorrente da anotação.

Para estruturar o texto, atribuímos relações RST às suas unidades. Como as relações são funcionais, o que importa é o tipo de efeito que elas podem produzir no leitor. Assim, elas podem ser descritas em termos das finalidades do escritor do texto, das suas suposições sobre o leitor e de determinados padrões proposicionais em relação ao conteúdo do texto. “As relações da estruturação do texto refletem as opções do escritor, de organização e apresentação; é nesse sentido que a RST é retórica” (Mann et al., 1992, p. 45)¹⁰.

Para atribuir relações às unidades, valemo-nos das definições das relações RST e observamos se a definição se aplica plausivelmente às unidades em questão. Um exemplo de definição de relação é o seguinte:

Nome da relação: EVIDENCE

Condições sobre núcleo (N): o leitor pode não acreditar em N em um grau satisfatório para o escritor do texto.

Condições sobre satélite (S): o leitor acredita na informação apresentada no S ou a achará crível.

Condições sobre combinação núcleo-satélite (N+S): a compreensão do conteúdo do S pelo leitor aumenta seu potencial de acreditar em N.

Efeito: a crença do leitor pela informação apresentada em N é aumentada.

Locus do efeito: N.

Cada campo de uma definição de relação especifica julgamentos particulares que os anotadores do texto devem fazer ao construir uma estrutura RST. Os anotadores têm acesso ao texto, têm conhecimento do contexto no qual ele se insere e compartilham as convenções culturais do escritor textual e dos leitores pretendidos, mas não têm acesso direto nem ao escritor do texto nem a outros leitores. Os anotadores têm acesso ao texto e podem ter conhecimento sobre o contexto no qual ele se insere, assim como podem compartilhar as convenções culturais do escritor ou dos leitores pretendidos. Porém, na maioria das vezes não têm acesso direto nem ao escritor do

¹⁰ Todas as transcrições de autores estrangeiros, neste relatório, apresentam tradução nossa.

texto nem a outros leitores supostos. Por isso, seus julgamentos devem ser de plausibilidade.

Na Figura 7, pode-se observar a análise RST de um trecho de texto diagramada na RSTTool, que ilustra a relação *evidence*, definida acima:

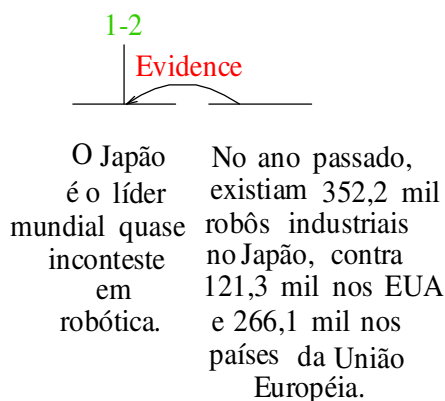


Figura 7. Ilustração da relação EVIDENCE

Na anotação do Corpus Summ-it, utilizamos a mesma classificação das relações utilizada em Pardo (2005), em relações intencionais e semânticas. A diferença entre esses tipos de relação não é tão trivial e várias teorias de discurso abordam-nos de modo distinto. Entretanto, um modo simples de entendê-la é pensar que as relações intencionais são aquelas que visam diretamente a alterar o comportamento ou crença do leitor em relação ao conteúdo proposicional apresentado pelo escritor (como a relação *evidence*) e, assim, tem uma força retórica mais “expressiva” ou uma conotação que visa a uma alteração explícita do comportamento do leitor. As semânticas, por sua vez, não teriam essa função, mas somente a de apresentar informações factuais, ou seja, são relações com função menos conotativa e mais informativa. Essas relações remetem a dependências inerentes dos conteúdos proposicionais ou do assunto abordado, propriamente dito, como a relação *cause*, que envolve unidades de informação que, naturalmente, caracterizam uma causa e um efeito. Um exemplo de senso comum para essa relação semântica pode ser dado pelo texto *Choveu. O chão ficou úmido.*

Para a anotação RST, consideramos também as mesmas relações estruturais que Pardo adotou e que foram propostas por Marcu (1997; 2000): *attribution*, *parenthetical* e *same-unit*. Essas relações, particularmente, fogem totalmente dos critérios originais da RST porque, simplesmente, referem-se à forma superficial do

inter-relacionamento das informações propositivas e nada têm a ver com seu caráter comunicativo, que é a tônica da RST. Por exemplo, a relação *parenthetical* indica informação adicional que foge ao fluxo natural do texto. Incluem-se aqui informações delimitadas por parênteses ou especificadas em notas de rodapé, por exemplo. A relação *same-unit* visa somente dar conta da existência de quebras de uma oração principal, ou do fluxo natural de informação já referido, mas agora pela introdução, no texto, de orações relativas restritivas ou de apostos.

Na RST, os critérios da classificação das relações têm como maior indicador o efeito que uma relação tem no leitor, definido para cada relação. Segundo Mann e Thompson (1988, p. 250):

Relações de conteúdo são aquelas cujo efeito pretendido é o de que o leitor *reconheça* a relação em questão; relações apresentativas são aquelas cujo efeito pretendido é *aumentar alguma inclinação* no leitor, como o desejo de agir ou o grau de atitude positiva, crença ou aceitação do núcleo.

Nessa definição, as relações de conteúdo são claramente as relações semânticas, enquanto as apresentativas são as relações intencionais. É o campo Efeito, numa definição de relação, que especifica o efeito pretendido pelo escritor sobre a reação do leitor ante um texto cuja relação interproposicional remete à relação RST correspondente. Assim, a funcionalidade do campo Efeito está intimamente relacionada à funcionalidade dos elementos constitutivos da mensagem, quais sejam, o núcleo e o satélite previstos na relação.

O campo Locus do efeito nos permite distinguir aquelas relações cujo *locus* está no núcleo e aquelas cujo *locus* está em ambos, no núcleo e no satélite. Quando o *locus* do efeito está no núcleo, como na relação de *evidence*, a nuclearidade pode representar a diferença qualitativa entre o essencial e o não essencial: o satélite seria a informação que dá apoio ao leitor, para a compreensão da informação apresentada no núcleo, mas não é essencial para esse entendimento ou aceitação. Quando o *locus* do efeito está em ambos, uma função diferente é prevista: a relação pode expressar características relativas entre núcleo e satélite que os fariam igualmente relevantes. O *locus*, neste caso, indicaria que toda informação, quer núcleo, quer satélite, remete ao foco do segmento textual.

Um exemplo de relação cujo *locus* do efeito está em ambos, no núcleo e no satélite, é dado pela relação *circumstance*. Sua definição é a seguinte (Mann et al., 1992):

Nome da relação: CIRCUMSTANCE

Condições sobre N: não há

Condições sobre S: apresenta uma situação já realizada

Condições sobre N+S: S apresenta o contexto sobre o assunto no qual o leitor deve interpretar a situação apresentada no N

Efeito: o leitor reconhece que a situação apresentada em S fornece o contexto para a interpretação de N

Locus do efeito: N e S

A definição dessa relação especifica que o satélite estabelece um contexto textual no qual o núcleo pode ser interpretado. Tendo o *locus* restrito a ambos, núcleo e satélite, a idéia é que o leitor deve focalizar ambas as informações para reconhecer que a situação apresentada no satélite fornece subsídios para a interpretação do núcleo.

Para efeito de identificação de informações essenciais, portanto, seria possível reconhecer, pelo campo Locus do efeito, quando a nuclearidade reflete o papel coadjuvante do satélite ou não. Quando ele aponta ambos, a nuclearidade refletiria o papel simbiótico do núcleo e do satélite, para o reconhecimento do leitor da relação proposicional entre núcleo e satélite. No entanto, não há um consenso sobre o uso desse campo na modelagem do discurso para fins computacionais, razão pela qual ele foi ignorado no Projeto ProCaCoSa em prol da praticidade: embora seja importante, para a anotação RST, distinguir esses papéis variados que a RST permite atribuir às relações, ao anotar os textos do Corpus Summ-it procuramos analisar e discutir as delimitações impostas semântica ou intencionalmente pelas definições da teoria diretamente no contexto de cada um dos segmentos proposicionais a inter-relacionar. Utilizamos, para isso, as definições das restrições de cada relação RST e de seu efeito, somente.

O conjunto de relações adotado foi aquele delimitado por Pardo com base em sua análise de um corpus de textos científicos do domínio da Ciência da Computação. Seu objetivo era o de implementar o analisador discursivo DiZer (Pardo et al., 2004; Pardo, 2005). Assim, os respectivos indicadores de relações RST incorporados ao sistema como seus únicos marcadores discursivos ou expressões indicativas do

discurso certamente são dependentes do corpus. Porém, ele é suficientemente abrangente, no que concerne ao conjunto de relações, para ser aplicável a outros domínios, razão para considerá-lo também neste trabalho, para a análise RST do corpus Summ-it.

Esse conjunto é composto por 32 relações: *antithesis, attribution, background, circumstance, comparison, concession, conclusion, condition, elaboration, enablement, evaluation, evidence, explanation, interpretation, justify, means, motivation, non-volitional cause, non-volitional result, otherwise, parenthetical, purpose, restatement, solutionhood, summary, volitional cause, volitional result, contrast, joint, list, same-unit, sequence*. Dessas, 24 fazem parte do conjunto original de Mann e Thompson (1987) e 8 fazem parte do conjunto proposto por Carlson e Marcu (2001) – *attribution, comparison, conclusion, explanation, means, parenthetical, list, same-unit*.

É importante salientar que, embora tenhamos adotado o mesmo conjunto de relações do DiZer, utilizamos suas definições apresentadas nos textos originais (Mann & Thompson, 1987; Carlson & Marcu, 2001), assim como aquelas presentes no sítio da RST (<http://www.sfu.ca/rst/01intro/definitions.html>).

A seguir apresentamos as etapas de anotação na ordem em que elas foram realizadas.

4.2. Segmentação proposicional

A segmentação proposicional consistiu em delimitar primeiramente as unidades textuais ou, segundo Carlson e Marcu (2001), as unidades discursivas elementares, já introduzidas neste relatório e identificadas por sua sigla, EDUs. Por ser considerada uma unidade ou proposição mínima de significado, uma EDU é sempre uma folha de uma árvore RST. Como já mencionado, o tamanho dessa unidade é arbitrário para a RST, podendo constituir-se de uma única oração, uma sentença ou mesmo de parágrafos inteiros ou unidades ainda maiores, desde que veiculem proposições completas, isto é, uma unidade de informação coerente. Porém, as

unidades devem ter integridade funcional independente e, uma vez determinada sua granularidade, todas as EDUs devem obedecer a esse mesmo padrão¹¹.

4.2.1. Segmentação proposicional do Corpus Summ-it

Foi adotada a mesma convenção de Carlson e Marcu para a segmentação de nosso corpus: eles sugerem que se considere a oração como a unidade elementar do discurso. Para delimitá-la, portanto, usamos indícios lexicais e sintáticos típicos de construções oracionais, os quais contribuem para a determinação das fronteiras proposicionais.

4.2.2. Problemas de Segmentação do Corpus Summ-it

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

- a) Consideramos orações reduzidas (adverbiais e relativas não-restritivas) como EDUs, como em

[1] O estudo, [2] feito por pesquisadores do Imperial College, em Londres, [3] mostra que (...).

em que [1] e [3] constituem uma única EDU, isolada da EDU [2].

- b) Não consideramos orações restritivas como EDUs, como em

[1] (...)os pesquisadores analisaram o fígado de mulheres que haviam sofrido um transplante de medula óssea (...).

- c) Mesmo que não oracionais, segmentos que estão entre parênteses, travessões ou outros delimitadores gráficos foram considerados EDUs e foram relacionados pela relação PARENTHETICAL, como nos exemplos abaixo:

[1] A projeção dos cientistas para o ano 2025 é que 3,3 bilhões de pessoas não tenham mais água para irrigação – [2] a atividade humana que mais consome o líquido.

¹¹ A granularidade, aqui, expressa o nível de complexidade de cada EDU. Se ela tiver granularidade sentencial, por exemplo, é possível que ela seja multi-proposicional. Entretanto, se sua granularidade for oracional, do ponto de vista da Linguística Textual, uma EDU seria, via-de-regra, mono-proposicional.

[1] A tecnologia difere do biodiesel utilizado em outras partes do mundo, que usa o metanol – [2] um derivado do petróleo.

A relação *parenthetical* possibilita que segmentos não oracionais sejam considerados EDUs, desde que sejam separados por delimitadores gráficos “fortes”. Para Marcu (2000), as vírgulas seriam delimitadores “fracos” (isto é, não suficientemente fortes para fazer de uma não-oração uma EDU). Porém, nem sempre os delimitadores gráficos indicam a relação *parenthetical*. É o caso de segmentos como: “Pessoas nascidas na China têm mais facilidade de se lembrar de um objeto quando o vêem pela segunda vez com o mesmo fundo que aparecia na primeira olhada – o que já não acontece com os americanos.” Nesse caso, o segmento após o travessão estabelece uma relação *contrast* com os segmentos anteriores. Assim, assumimos que, na possibilidade de ser *parenthetical*, analisamos se não há outra relação mais proeminente que ela, a qual irá prevalecer na estrutura RST.

Por outro lado, alguns segmentos que não estão separados por delimitadores gráficos “fortes” podem estabelecer uma relação *parenthetical*. É o caso do exemplo dado em (a) acima, em que o segmento [2], que está entre vírgulas, se encaixa como satélite na descrição da relação *parenthetical*: “S apresenta informação extra relacionada ou complementar a N; S não pertence ao fluxo principal do texto”.

d) Mesmo que não oracionais, segmentos que indicam a autoria de discursos diretos ou indiretos são considerados EDUs e relacionados pela relação *attribution*. Isso implica assumir que opiniões de autoria acompanhadas da indicação do autor, quer direta (em transcrições literais), quer indiretamente (como no uso de “segundo fulano...”), são particionadas em duas EDUs. Carlson e Marcu (2001) distinguem esses dois casos denominando-os de *reported speech* e *reporting speech*, respectivamente, como nos casos abaixo:

[1] “É uma descoberta e tanto”, [2] disse o psicólogo César Ades.

[1] Segundo Kellner, [2] apesar de o animal ser um baixinho, (...).

[1] O ministro da Agricultura, Roberto Rodrigues, afirmou que [2] o uso da soja transgênica “é uma boa idéia”.

Além dessas diretrizes específicas adotadas para a segmentação oracional, também foi decidido consensualmente como lidar com outros aspectos dos textos do corpus,

como títulos, por exemplo. Como o corpus é composto de textos de divulgação científica publicados em contexto midiático, sua grande maioria apresenta, além dos títulos, subtítulos que delimitam porções de texto. Decidimos desconsiderar tanto os títulos quanto os subtítulos, suprimindo-os ao colocar os textos na RSTTool. Porém, registramos cada caso de supressão, para que houvesse um controle da maneira como foram tratados os textos do corpus.

Após delimitar todas as EDUs de um texto, o passo seguinte consistiu em estabelecer suas relações RST, ou seja, em construir sua estrutura retórica, o que foi reproduzido para cada texto do corpus Summ-it.

4.3. Estruturação retórica

4.3.1. Pressupostos teóricos e práticos

Para construir a estrutura hierárquica dos textos do corpus na RSTTool, valemo-nos dos procedimentos descritos nas obras principais de base: a de fundamentação teórica (Mann & Thompson, 1987), a de anotação prática (Carlson & Marcu, 2001) e a aplicada especialmente ao português (Pardo, 2005). Para cada texto, após sua segmentação, foram atribuídas relações RST a pares de EDUs nucleares e satélites indicados no texto como segmentos diretamente inter-relacionados. As relações entre essas EDUs, por delinearem segmentos maiores, resultaram em sub-árvores RST cuja raiz é uma relação RST e cujas folhas, as EDUs em foco. Essas sub-árvores, por sua vez, foram relacionadas a outras sub-árvores delineadas de forma análoga, formando segmentos maiores ainda. Sucessivamente, a constituição da estrutura hierárquica completa do texto se fez possível. Para textos coesos, é pressuposto da teoria que todas as EDUs se inter-relacionarão em uma única árvore RST, não sendo possível haver EDUs isoladas ou não conectadas.

A estrutura hierárquica do texto **CIENCIA_2000_17109**, reproduzido abaixo e extraído do corpus Summ-it, pode ser observada na Figura 8. Sub-árvores não ilustradas em detalhe são compactadas e indicadas por triângulos nessa figura.

Texto CIENCIA_2000_17109

[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.

[3] Células-tronco são células não-especializadas, capazes de dar origem a qualquer tipo de tecido. [4] As da medula óssea dão origem a células sanguíneas. [5] O estudo, [6] feito por pesquisadores do Imperial College, em Londres, [7] mostra que, além disso, elas são capazes de originar outro tipo de célula – [8] células hepáticas – [9] dentro do organismo humano.

[10] A descoberta possibilitará que pessoas com dano no fígado usem as próprias células-tronco [11] para produzir células hepáticas.

[12] "No futuro, [13] quando a produção de tecido hepático se tornar uma realidade, [14] o número de transplantes poderá ser minimizado", [15] disse à Folha por e-mail Joe Jackson, um dos autores do estudo que sai hoje na revista "Nature".

[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, [19] cujo doador havia sido um homem.

[20] A análise do DNA dessas células mostrou que elas continham o cromossomo Y, [21] encontrado apenas em células masculinas. [22] Isso indica que, de alguma forma, as células-tronco da medula óssea haviam sido capazes de "colonizar" o fígado das mulheres transplantadas. [23] O resultado deverá contribuir para o desenvolvimento de tecidos humanos para uso terapêutico, [24] dizem os autores.

É possível observar aqui que as relações são atribuídas recursivamente aos diversos níveis hierárquicos do texto. A relação *elaboration*, por exemplo, estabelece-se no nível elementar das EDUs entre os segmentos 18 e 19 ou entre os segmentos 20 e 21, por exemplo. Também ocorre em nível mais complexo do texto, como entre os segmentos 1-15 e 16-22.

4.3.2. Problemas de Estruturação do Corpus Summ-it

- **CIENCIA_2000_17108:** houve dúvida sobre como relacionar os segmentos seguintes: “A exploração alheia não tem limites. Nem mesmo no reino animal.” O segmento sublinhado, considerado satélite pelos anotadores, é algo que intensifica o segmento anterior. Porém, não há, no conjunto de relações adotado, nenhuma relação que pareça desempenhar essa função de intensificação. Em

casos como esse, com base em sugestões recebidas¹², convencionamos utilizar a relação *elaboration*, que, por sua definição, é capaz de abranger casos como esse (Figura 9). Vale notar que essa estratégia de adotar a relação *elaboration* em casos obscuros também é adotada por outros autores, como o faz Pardo (2005).

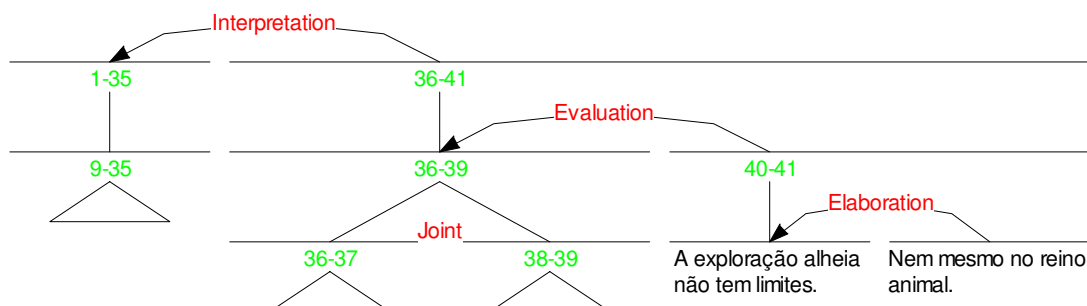


Figura 9. ELABORATION entre EDUs do texto CIENCIA_2000_17108

- Mesmo texto **CIENCIA_2000_17108**: houve dúvida sobre como relacionar os segmentos: “[A larva passa de 7 a 14 dias ali dentro,]₁ [fartando-se do sangue do aracnídeo,]₂ [até estar madura o suficiente.]₃”. Cogitamos o uso das relações *explanation*, *circumstance* ou *condition* entre as EDUs 1 e 3, mas concluímos pela relação *non-volitional result*, pois, ao final das discussões, decidimos que a maturação (EDU3) – compreendida como um processo natural de desenvolvimento do organismo em questão – é um resultado involuntário do ciclo natural do animal (EDU1) e que casos análogos a este deveriam ser tratados como *non-volitional result*. Essa escolha se deveu à rigidez da forma de estruturação RST de um texto, que é decorrente das próprias definições das relações entre núcleo e satélite. Isto configura um problema adicional à decisão sobre múltiplas relações: qualquer uma delas deveria se estabelecer entre as EDUs 2 e 3 e não entre as EDUs 1 e 2, já que a progressão que resulta no evento de amadurecimento (EDU3) se dá pelo evento de fartar-se de sangue (EDU2) e não de passar tantos dias ali dentro (EDU1). Isso fica evidente na forma gráfica da Figura 10. Uma dúvida que surge aqui é, então, porque não fazer a EDU3 ser satélite da EDU2, pela relação *non-volitional result* (Figura 11). Claramente, essa opção também não é satisfatória, pois a EDU1 assume papel importante para a progressão do evento até sua interrupção (marcada pela

¹² Sobre essa e outras dúvidas, elaboramos um relatório de dúvidas e consultamos outros especialistas em RST, particularmente, Thiago A.S. Pardo e Eloize M. Seno.

partícula “até”). Para precisar melhor essa relação de *non-volitional result* entre as EDUs 2 e 3 seria necessário, talvez, criar uma outra definição de resultado que incorporasse a idéia de progressão, situação não prevista no elenco de relações RST em uso.

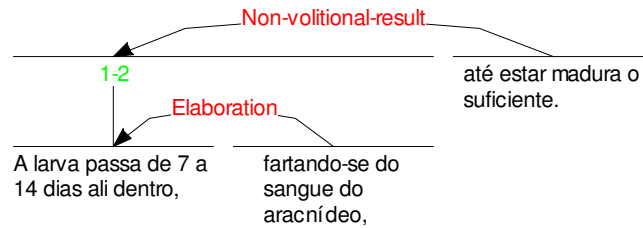


Figura 10. NON-VOL. RESULT entre EDUs do texto CIENCIA_2000_17108

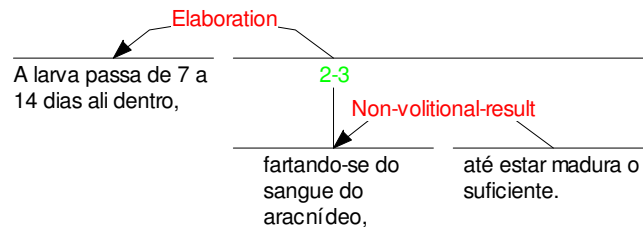


Figura 11. Alteração da estrutura ilustrada na Figura 10

- **CIENCIA_2000_17101:** “[O debate surgiu após estudos em Ruanda e na Tailândia,]₁ [em que cientistas deram a grávidas com HIV um regime de AZT mais breve do que o recomendado.]₂ [Queriam saber se o regime curto era melhor que nada]₃ [para impedir a contaminação do feto.]₄”. O segmento 3-4 apresenta uma informação que indica por que os cientistas executaram uma determinada ação. Surgiu, então, a hipótese de esse segmento 3-4 caracterizar motivação para os cientistas haverem dado a grávidas o regime de AZT (EDU2), o que levaria à relação *motivation*. Porém, essa relação, por sua definição, não poderia ser aplicada nesse caso, já que ela envolve explicitamente uma motivação (no seu satélite) para que o leitor realize uma ação. Claramente, a escolha dessa relação levaria a um erro de análise RST, porém, reiteramos a validade de interpretar a relação entre a EDU2 e o segmento 3-4 como uma relação de motivação. As alternativas possíveis, nesse caso, foram *justify* e *explanation*. Como a primeira é intencional e a segunda é semântica, optamos pela relação *explanation* (muito embora essa ambigüidade possa permanecer, devido à sutileza entre a classificação de relações como intencionais ou semânticas), como ilustramos na Figura 12. Esse exemplo é interessante para se reafirmar o problema do caso anterior: enquanto gostaríamos de associar mais diretamente a relação *non-volitional result* entre as

EDUs 2 e 3 (entre EDU3 e satélite do segmento 1-2, mais exatamente), como mostra a Figura 11, neste caso a relação *explanation* se estabelece naturalmente entre os núcleos dos dois segmentos envolvidos (1-2 e 3-4). Isso só vem indicar novamente que as definições RST se aplicam sem conflitos para alguns casos, enquanto não são suficientes para dar conta de outros.

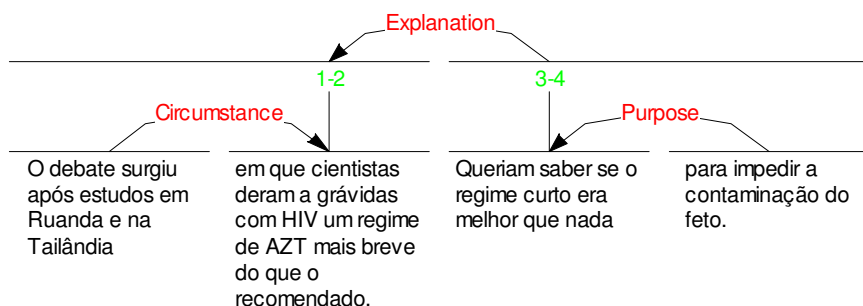


Figura 12. EXPLANATION entre 1-2 e 3-4 do texto CIENCIA_2000_17101

- **CIENCIA_2005_28756:** Uma dúvida muito freqüente foi sobre que relação marcar para casos como o que ocorre em “[Se essa hipótese for verdadeira,]₁ [os pesquisadores já sabem até em que grupo marsupial encaixar o caco:]₂ [ele pertenceria ao grupo dos polidolopimorfos, comedores de frutas parecidos com gambás ou cuícas que hoje estão extintos.]₃” O segmento 1-2 anuncia algo, o que está apresentado na EDU3. Esta, por sua vez, apenas explicita o que foi anunciado antes. Também por sugestão de Thiago Pardo, convencionamos utilizar a relação *attribution* (Figura 13), já que sua definição (dada abaixo) deixa evidente que uma proposição é de autoria de alguém – nesse caso, dos pesquisadores.

Nome da relação: ATTRIBUTION

Condições sobre núcleo (N): N apresenta uma expressão, fala ou pensamento de alguém ou algo

Condições sobre satélite (S): S apresenta alguém ou algo que produz N

Condições sobre combinação núcleo-satélite (N+S): S e N indicam, respectivamente, a fonte de uma mensagem e a mensagem, propriamente dita

Efeito: o leitor é informado sobre a mensagem e sobre quem ou o que a produziu

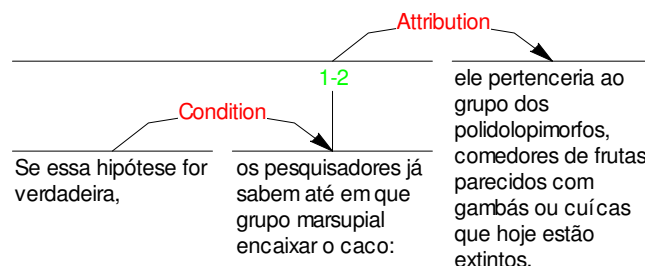


Figura 13. ATTRIBUTION entre 1-2 e EDU3 do texto CIENCIA_2005_28756

Esses exemplos ilustram problemas de determinação de relações RST entre segmentos textuais, que foram resolvidos pela consulta às suas definições e à verificação de sua plausibilidade para os segmentos em foco. Ilustram, também, que as definições não são sempre suficientes, havendo discordância e obscuridade para a anotação em vários casos, só podendo ser sanadas com discussão e busca de consenso. Como resultado, definimos consensualmente as diretrizes aplicadas à anotação do Corpus Summ-it.

A seguir, são comentados os resultados encontrados na análise dos 12 textos do Corpus Summ-it, que podem mostrar algumas regularidades em termos das relações e das estruturas RST sobre o gênero de texto analisado.

5. Resultados da análise do corpus parcial

A incidência das relações encontradas na amostra de 12 textos do Corpus Summ-it pode ser observada na Tabela 2. É dada a quantidade absoluta de cada relação, assim como sua representatividade no corpus.

Tabela 2. Incidência das relações RST nos 12 textos do corpus Summ-it

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

Algumas dessas relações foram encontradas principalmente no nível elementar, isto é, entre EDUs. Além disso, essas relações, em sua maioria, apresentaram marcadores textuais que as identificavam. A Tabela 3 apresenta a incidência dessas relações que apareceram no nível elementar dos textos e de todas as relações, dentre essas, que foram identificadas explicitamente por marcadores textuais. São representadas aqui somente aquelas para as quais discriminamos os marcadores textuais (391) em vez das 400 apresentadas na Tabela 2. Os números que aparecem entre parênteses junto ao nome de cada relação reproduzem o total de ocorrências no corpus, apresentado na Tabela 2. As relações marcadas com * são multinucleares. Nas relações mononucleares, os marcadores aparecem sempre nos satélites, ou no seu início ou no meio. Nas relações multinucleares, eles aparecem em um dos núcleos. Na relação *same-unit*, os marcadores são agregados ao satélite que se interpôs entre os núcleos da mesma unidade.

Tabela 3. Incidência de relações no corpus

Relações	# Ocorrências	# Oc. marcadas	Marcadores textuais
Elaboration (93)	50	36	<u>pronomes relativos</u> (que-13, onde-4, cujo-2, como-1, em que-1); <u>verbos no particípio</u> -9 (assinadas, encontrado, aprimorados, todas causadas, escolhido, realizado e promovido, desenvolvido, ocultas, encerrado); <u>verbos no gerúndio</u> -4 (forçando, fartando-se, prestando atenção, gesticulando); <u>advérbios</u> -2 (anualmente, literalmente)
Attribution (50)	50	48	<u>verbos</u> (disse, diz, dizem-20, afirmou, afirma-14, conta, contou-2, sugerem-2, explicou-1, argumentaram-1, descobriu-1, defendeu-1, resume-1); <u>conjunções</u> (segundo-4); <u>preposições</u> (para-1)
Patenthetical (42)	42	42	<u>parênteses</u> -27; <u>travessões</u> -10; <u>colchetes</u> -4; <u>verbo no particípio</u> (feito-1)
Same-Unit* (34)	34	31	<u>parênteses</u> -16; <u>travessões</u> -4; <u>pronomes relativos</u> (que-3); <u>conjunções</u> (segundo-2, quando-1); <u>verbos</u> (disse-1, forçando-1, desenvolvido-1, feito-1, aprimorados-1)
Purpose (20)	20	20	<u>preposições</u> (para-18); <u>conjunções</u> (com o objetivo de-1, na tentativa de-1)
Circumstance (15)	12	11	<u>conjunções</u> (quando-4; sem-2; ao-2; enquanto-1; assim que-1); <u>pronome relativo</u> (onde-1)
List* (17)	10	10	<u>conjunções</u> (e-10)
Concession (8)	8	8	<u>conjunções</u> (mas-6, apesar de-2)
Condition (9)	7	7	<u>conjunções</u> (se-5); <u>verbos no particípio</u> (não satisfeita-1, dopada-1)
Evaluation (20)	7	5	<u>adjetivos</u> (menos precisos-1; minúsculas-1, incompleto-1); <u>verbos</u> (não pode-1, parece até-1)
Background (19)	4	1	<u>verbo no particípio</u> (batizado-1)
Non-Volitional Result (5)	4	4	<u>verbos no gerúndio</u> (aumentando-1, matando-a-1, destruindo-1); <u>preposição</u> (até-1)
Contrast* (10)	4	2	<u>conjunções</u> (enquanto-2)
Otherwise (3)	3	3	<u>conjunções</u> (em vez de-3)
Interpretation (26)	3	2	<u>expressões</u> (é como se-1, isso indica que-1)
Antithesis (2)	2	2	<u>conjunções</u> (mas-2)
Means (2)	2	2	<u>verbos no gerúndio</u> (olhando-1, usando-1)
Non-Volitional Cause (2)	2	2	<u>preposição</u> (por-1); <u>expressão</u> (se deve a-1)
Sequence* (7)	1	1	<u>conjunção</u> (e-1)
Conclusion (4)	1	0	
Justify (1)	1	0	
Explanation (2)	1	0	
TOTAIS de relações			
391	268	237	

Conforme mostra essa tabela, 69% das relações nos 12 textos analisados apareceram no nível elementar dos textos; destas, 88,4% apresentaram marcadores

textuais. É possível observar que há relações que apresentam todas as suas ocorrências (ou a maioria delas) no nível elementar. É o caso das relações *attribution*, *parenthetical*, *same-unit*, *purpose*, *circumstance*, *concession*, *condition*, *non-volitional result*, *otherwise*, *antithesis*, *means*, *non-volitional cause* e *justify*. Por outro lado, há relações que dificilmente aparecem nesse nível. É o caso das relações *evaluation*, *background*, *contrast*, *interpretation*, *sequence* e *conclusion*. Essas relações foram encontradas quase que exclusivamente em níveis macro-estruturais dos textos. Além disso, elas quase não apresentaram marcadores textuais que as identificassem. Um caso ímpar é o da relação *elaboration*, que, por sua versatilidade, aparece tanto no nível elementar quanto no nível macro-estrutural dos textos.

A incidência de determinadas relações nos níveis macro-estruturais analisados está ligada à superestrutura do gênero de texto em questão (texto de divulgação científica publicado em contexto midiático). O propósito desse gênero textual é divulgar uma pesquisa a um público variado. Portanto, alguns elementos textuais desse gênero são tipificados em sua superestrutura: (i) a menção à pesquisa divulgada; (ii) a apresentação dos procedimentos metodológicos utilizados na pesquisa divulgada; (iii) a avaliação e a interpretação dos pesquisadores ou de outros membros da comunidade científica sobre a repercussão da pesquisa.

Esses elementos superestruturais foram observados em todos os textos analisados. Para cada um deles, adotamos convenções particulares para atribuir algumas das relações RST. A menção à pesquisa divulgada é sempre o segmento nuclear mais saliente na estrutura hierárquica dos textos, já que a finalidade do gênero textual em questão é divulgar alguma pesquisa. Todas as relações macro-estruturais de cada texto estão ligadas como satélites a esse segmento nuclear. É o que pode ser visto na Figura 14, entre os segmentos 1-2 e o resto do texto:

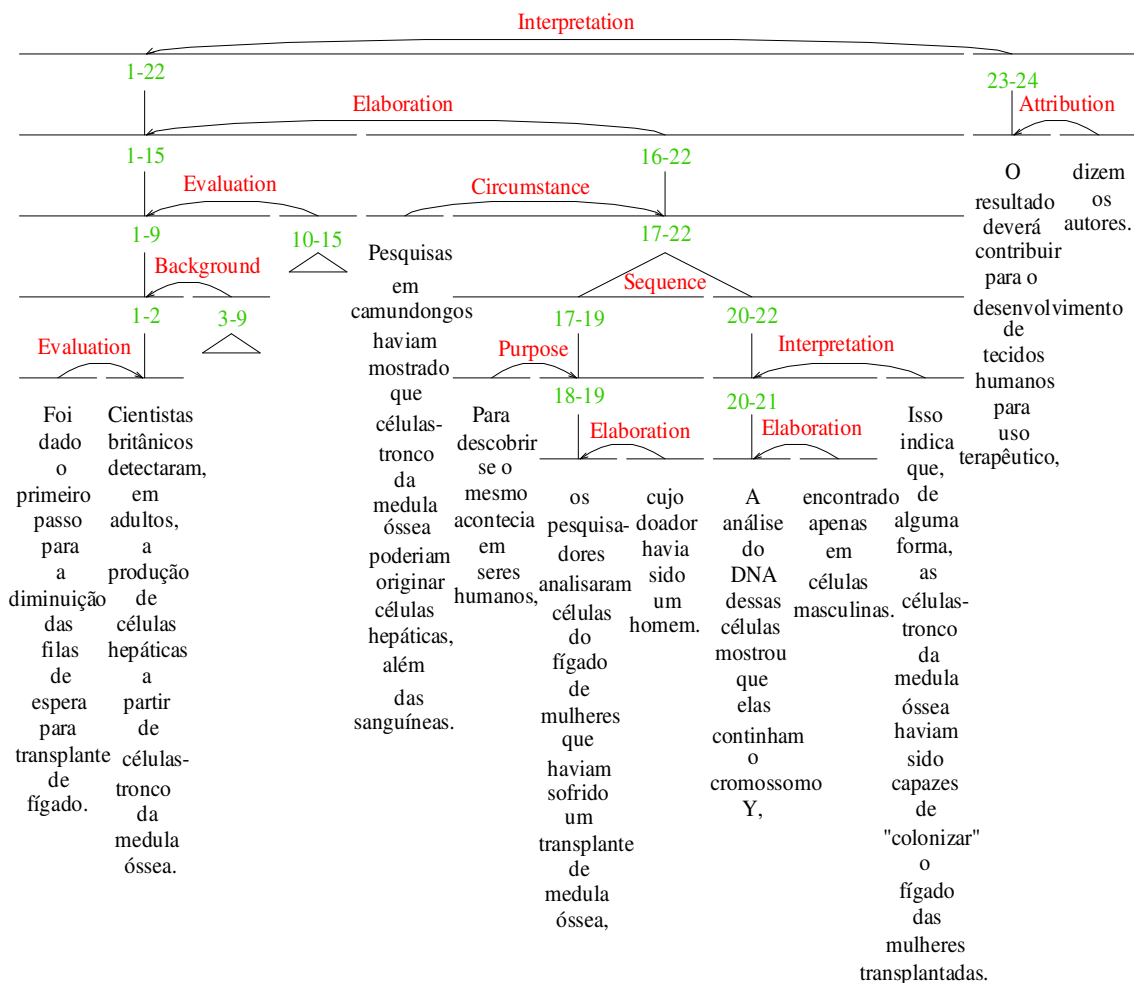


Figura 14: Análise do texto CIENCIA_2000_17109

A apresentação dos procedimentos metodológicos utilizados na pesquisa divulgada (ou os procedimentos observados nos objetos pesquisados) geralmente constitui um segmento macro-estrutural diretamente ligado ao segmento nuclear mais relevante na estrutura hierárquica do texto. A relação observada nesse tipo de segmento é *elaboration*, ocorrendo em 58% dos 12 textos analisados. Um exemplo de ocorrência dessa relação no nível macro-estrutural está entre os segmentos 1-15 e 16-22 na figura acima.

A interpretação e a avaliação dos pesquisadores ou de outros membros da comunidade científica sobre a repercussão da pesquisa também constitui um segmento macro-estrutural, como no caso anterior. Em geral, elas se sobrepõem à apresentação dos procedimentos metodológicos. Isso faz sentido, se considerada a prática de pesquisa: avaliações ou interpretações são apresentadas, via de regra, após a descrição da metodologia adotada (cf. p.ex., Hoey, 1983; Jordan, 1980; Weissberg & Buker, 1990). A relação *interpretation* ocorreu em 33% dos textos

analisados, enquanto a relação *evaluation*, em 25% dos textos. Um exemplo da primeira relação é apresentado na figura acima entre os segmentos 1-22 e 23-24; um exemplo da segunda pode ser observado entre os segmentos 1-4 e 5-16 da Figura 15:

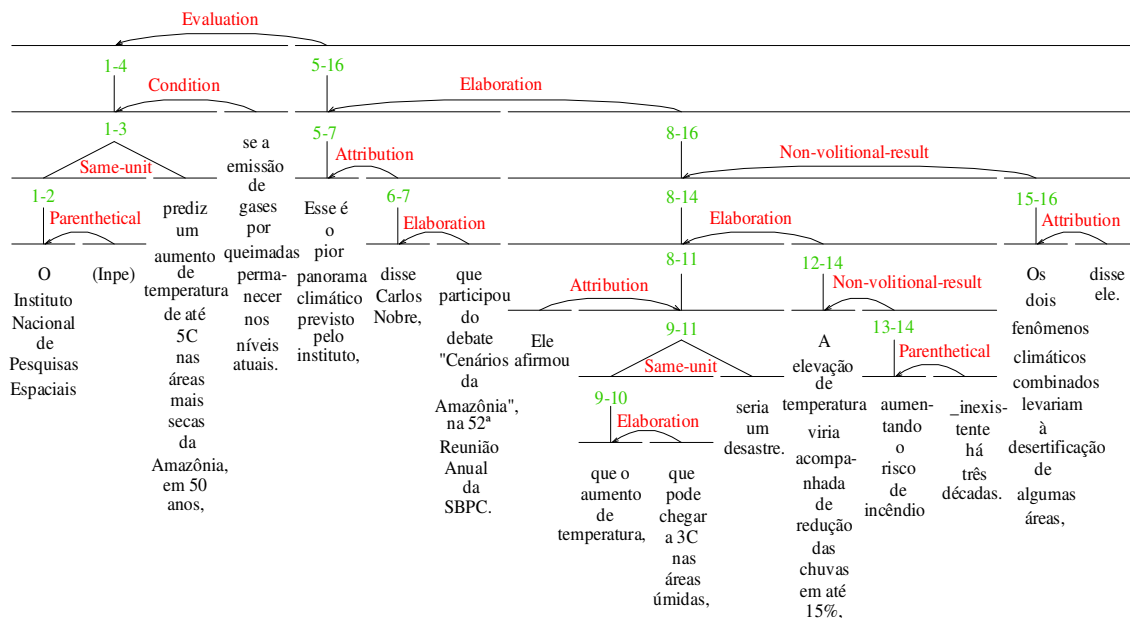


Figura 15. Análise de parte do texto CIENCIA_2000_17082

Além desses elementos superestruturais, que caracterizam a organização com base no gênero textual em questão e restringem as possibilidades de macro-estruturação dos textos, há outros elementos indicativos da macro-estrutura que também se mostraram freqüentes nos textos analisados. Por exemplo, 41,6% dos textos analisados apresentaram, no seu início, frases e expressões que parecem estabelecer uma proximidade com o leitor. Depreendemos que isso acontece porque, em textos de divulgação científica publicados em contexto midiático, diferentemente do que acontece no texto científico dirigido aos pares (publicado em periódicos científicos de uma dada comunidade científica), não se espera que o leitor esteja interessado de antemão na pesquisa que será veiculada. Além disso, nos meios de comunicação de alta circulação – como jornais e revistas – a matéria é um produto a ser comercializado. Portanto, o texto de divulgação científica publicado em contexto midiático deverá, antes de tudo, captar a atenção do leitor.

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

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

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

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

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

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

Essa função de captação do leitor é pragmática. Portanto, os trechos de texto exemplificados seriam satélites de alguma relação intencional (ou apresentativa), ou seja, de uma relação cujo efeito fosse aumentar alguma inclinação no leitor para a leitura do núcleo (e cujo *locus* do efeito estivesse apenas no núcleo). Pensamos primeiramente na relação *preparation*¹³, que, como se pode observar pela sua descrição (abaixo) e pelo seu uso mais recente por Mann (2006)¹⁴, seria aplicável a esses trechos.

Nome da relação: PREPARATION

Condições sobre núcleo (N): Não há

Condições sobre satélite (S): Não há

Condições sobre combinação núcleo-satélite (N+S): S precede N no texto e tende a compelir o leitor a se interessar por ler a informação em N

Efeito: o leitor se mostra disposto a ler o conteúdo de N

Essa relação, porém, não figura entre as relações do conjunto adotado neste trabalho e não há nenhuma relação intencional cuja descrição pareça se aplicar aos trechos em questão. Assim, optamos pelo uso das relações semânticas *evaluation*¹⁵ e *interpretation*¹⁶. Essas relações, embora semânticas, apresentam na sua descrição elementos que evidenciam a posição do escritor do texto, no satélite, frente ao que é dito no núcleo.

¹³ Efeito: o leitor se sente mais preparado, interessado ou orientado para ler o núcleo.

¹⁴ Mann, William. Relation definitions. <http://www.sfu.ca/rst/01intro/definitions.html> (3 abr 2006).

¹⁵ Efeito: o leitor reconhece que a informação em S serve para avaliar a situação em N e reconhece o valor que lhe é atribuído.

¹⁶ Efeito: o leitor reconhece que a informação em S relaciona N a um contexto de idéias não envolvidas diretamente com o conhecimento apresentado em N.

Finalmente, um último caso de organização estereotipada no nível macro-estrutural que, do ponto de vista de estruturação RST, parece nos obrigar a determinar uma relação para manter a coesão estrutural, mesmo que, semanticamente, não pareça que ela se aplique ao contexto, remete usualmente aos segmentos finais dos textos, nesse gênero estudado. É o caso de intervenções que parecem indicar a necessidade do autor de concluir seu texto com alguma sentença, mesmo que, do ponto de vista informativo, seu conteúdo fuja àquele expresso no contexto global do texto. Por exemplo, no texto **CIENCIA_2005_28747**, em que a informação mais nuclear é expressa pelo segmento 3-6 – “Pesquisadores da Universidade de Michigan em Ann Arbor, nos Estados Unidos, sugerem que os olhos de asiáticos tendem a ver uma imagem no seu conjunto, prestando tanto atenção ao que está em primeiro plano quanto no fundo, enquanto os americanos demoram mais o olhar no objeto central de um quadro.” – a EDU32, que corresponde à sentença “O estudo está na revista ‘PNAS’”, não se relaciona semanticamente ao núcleo. No entanto, como um texto deve ter uma estrutura RST inteiramente coesa, essa EDU tem que ser interconectada ao restante. Adotamos, aqui, a relação *elaboration* (Figura 16), lembrando que esta é uma relação adotada no caso geral, por exemplo, quando nenhuma outra parece se aplicar, sobretudo nos sistemas automáticos (Pardo, 2005).

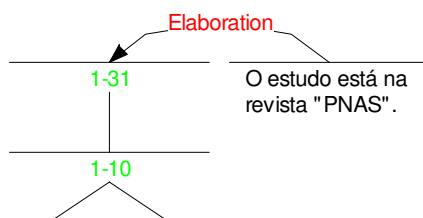


Figura 16. Finalização estrutural do texto CIENCIA_2005_28747

6. Considerações finais

A tarefa de anotação de corpus com estruturas retóricas implica um conjunto de procedimentos que permitem – e também impõem – aos analistas reflexões não apenas referentes aos problemas e desafios relativos à RST. A manipulação de textos com vistas à identificação e marcação das relações retóricas utilizadas demanda uma análise que envolve, pela natureza do objeto, questões de textualidade (coesão e coerência), de gêneros textuais e mesmo de pragmática.

A partir da experiência com os 12 textos cuja análise e estruturação reportamos neste relatório, pudemos derivar importantes considerações com relação aos padrões

de utilização de relações retóricas para o gênero contemplado pelo corpus (jornalístico de divulgação científica), identificando o uso preponderante de marcadores no nível elementar (microestrutural) do texto e enumerando-os.

Os dados apresentados aqui servem de ponto de partida para trabalhos futuros envolvendo o restante do Corpus Summ-it, assim como para análises RST de outros corpora. No momento, há a perspectiva de reproduzir os resultados obtidos com a análise dos 12 textos no restante do corpus (38 textos). Outros padrões retóricos também deverão ser explorados para o mesmo gênero. Além disso, as decisões consensuais na anotação RST parcial do Corpus Summ-it apresentadas aqui deve ser usada para complementar decisões de práticas anteriores, sobretudo para o português – representadas significativamente pelos trabalhos de Thiago Pardo (2005) e Eloize Seno (2005) – consolidando e elucidando os critérios de anotação. Esses desdobramentos, enfim, devem visar uma maior consistência nas tarefas futuras de anotação RST.

Referências bibliográficas

- Bonini, A. (2001). Gênero Textual como Signo Lingüístico: Os Reflexos da Tese da Arbitrariedade. *Linguagem em (Dis)curso*. 1(2): 123-135.
- Bronckart, J.P. (1999). Atividade de Linguagem, Textos e Discursos: Por um Interacionismo Sócio-Discursivo. Tradução de Anna Rachel Machado. Educ. São Paulo.
- Carlson, L.; Marcu, D. (2001). *Discourse Tagging Reference Manual*. ISI Technical Report ISI-TR-545.
- Cristea, D.; Ide, N.; Romary, L. (1998). Veins Theory: A Model of Global Discourse Cohesion and Coherence. *In the Proceedings of the Coling/ACL' 1998*, pp.281-285. Montreal, Canadá.
- De Beaugrande, R; Dressler, W. U. (1981). *Introduction to Text Linguistics*. Longman. New York.
- Antonio, J.D. (2004). *Estrutura Retórica e Articulação de Orações em Narrativas Orais e em Narrativas Escritas do Português*. Tese de doutorado. UNESP, Araraquara.
- Halliday, M. A.K.I; Hasan, R. (1976). *Cohesion in English*. Longman. London.
- Hoey, M. (1983). *On the Surface of Discourse*. George Allen & Unwin (Publishers) Ltd.
- Jordan, M.P. (1980). Short Texts to Explain Problem-Solution Structures – and Vice Versa. *Instructional Science* 9, pp. 221-252.
- Weissberg, R.; Buker, S. (1990). *Writing up Research: Experimental Research Report Writing for Students of English*. Prentice Hall, Inc.
- Koch, I. G. V. (2004). *A coesão textual*. Contexto Editora. São Paulo.
- Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

- Mann, W.C.; Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.
- Mann, W.C.; Matthiessen, C.; Thompson, S.A. (1992). Rhetorical structure theory and text analysis. In William C. Mann and Sandra A. Thompson, (eds.), *Discourse Description: Diverse Linguistic Analyses of a Fund-raising Text*, pp. 39-78. John Benjamins, Amsterdam/Philadelphia.
- Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD Thesis, Department of Computer Science, University of Toronto.
- Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. Cambridge, Massachusetts.
- Marcuschi, L. A. (1983). *Lingüística de texto: como é e o que se faz*. Universidade Federal de Pernambuco, Série Debates 1. Recife, PE.
- O'Donnell, M. (1997). RSTTool: An RST Analysis Tool. In *Proc. of the 6th European Workshop on Natural Language Generation*, Gerhard-Mercator University, Duisburg, Alemanha.
- Pardo, T.A.S. (2005). *Métodos para Análise Discursiva Automática*. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Junho, 211p.
- Pardo, T.A.S.; Nunes, M.G.V.; Rino, L.H.M. (2004). DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. In *the Proceedings of the 17th Brazilian Symposium on Artificial Intelligence – SBIA (Lecture Notes in Artificial Intelligence 3171)*, pp. 224-234. São Luis-MA, Brazil. September, 29 - October, 1.
- Sparck Jones, K. (1999). Automatic Summarizing: factors and directions. In I. Mani and M. Maybury (eds.), *Advances in Automatic Text Summarization*, pp. 1-12. The MIT Press. Cambridge.
- Sporleder, C., & Lascarides, A. (2005). Exploiting linguistic cues to classify rhetorical relations. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 532-539. Borovets, Bulgaria.
- Seno, E.R.M. (2005). *Especificação de Heurísticas de Sumarização de Estruturas RST com Base na Preservação dos Elos Co-Referenciais*. Dissertação de Mestrado. Departamento de Computação, UFSCar.
- Swales, J. (1992). *Genre Analysis*. Cambridge University Press. Cambridge.
- Van Dijk, T.A. (1979). Recalling and Summarizing Complex Discourse. In Burghart, W. and Hölker, K. (eds.), *Text Processing Textverarbeitung*. Walter de Gruyter. Berlin.

Anexo A – Textos originais do Corpus Summ-it

1. 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).

(WILSON SILVEIRA)

2. 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 Tyrannosaurus 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).

3. 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.

4. 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.

(CLAUDIO ANGELO)

5. 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.

6. CIENCIA_2000_17112

O mundo está mais seco do que se imaginava. Um estudo publicado na edição de hoje da revista "Science" afirma que 1,75 bilhão de pessoas já enfrentam severa escassez de água no planeta.

A estimativa anterior, da ONU (Organização das Nações Unidas), calculava em meio bilhão o número de indivíduos expostos atualmente ao problema.

Por severa escassez de água potável, entende-se, segundo a ONU, o uso de mais de 40% das reservas do líquido disponíveis em uma região para consumo industrial, doméstico e agrícola.

O trabalho foi coordenado pelo geocientista Charles Vörösmarty, da Universidade de New Hampshire, nos Estados Unidos.

Para realizar o cálculo, a equipe de Vörösmarty dividiu o mundo em 60 mil microrregiões. Depois, estimou a quantidade de água doce sustentável (presente em rios e reservatórios de superfície) disponível em cada região.

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

Todos os cálculos anteriores levavam em conta macrorregiões, como países e continentes. Eram, portanto, menos precisos.

"O que nós fizemos foi um ajuste fino", disse o pesquisador à Folha, por telefone. "Descobrimos que os recursos hídricos locais em algumas áreas estão simplesmente esgotados", afirmou.

Metrópoles na mira

As áreas mais atingidas, claro, são as regiões áridas do norte da África, da Ásia Central e do Oriente Médio. Mas zonas de intensa urbanização recente, como o sul dos EUA e o norte do México, também foram incluídas no novo mapa da escassez.

"A demanda aumenta de forma drástica no mundo todo", afirmou o especialista em recursos hídricos José Galizia Tundisi, do Instituto Internacional de Ecologia, em São Carlos (SP).

"As metrópoles não têm recursos hídricos suficientes para suportar o crescimento populacional", disse Tundisi.

7. CIENCIA_2000_17113

Um tratamento para a obesidade que faz você comer mais, perder peso e reduzir a taxa de gordura do corpo é o que sugere um estudo britânico publicado hoje na revista científica "Nature".

Por enquanto é só sugestão: o tratamento foi testado em camundongos. Mas os resultados levaram os cientistas a chamar o próprio estudo de "abordagem promissora" contra a obesidade.

No centro do método está um gene humano descoberto recentemente, o UCP-3, cujos mecanismos de ação ainda não são totalmente conhecidos. Sabe-se que está envolvido no processamento de energia pelas células e que um gene da mesma família, o UCP-1, está ligado à queima de gordura.

O gene UCP-3 foi inserido em camundongos e manipulado para produzir, em excesso, a proteína determinada por ele.

Os camundongos com essa alteração genética comeram até 54% mais que os camundongos normais. Apesar disso, pesavam até 23% a menos que seus companheiros. A porcentagem de tecido adiposo (gordura) sobre o volume total do corpo dos bichos também diminuiu _nos machos, em 44%; nas fêmeas, em 57%.

Sua atividade física não diferiu significativamente em relação à dos camundongos normais. Isso quer dizer que os camundongos transgênicos reduziram a gordura de seu corpo, em relação à massa muscular, sem fazer ginástica.

Os animais magros consumiram mais energia até para respirar. Em vez de armazenar a comida como gordura, transformaram-na em calor.

Novos remédios

O fato de o UCP-3 ser um gene humano facilita aplicações clínicas da pesquisa. Um dos caminhos seria sua superestimulação.

"Esse é um alvo viável para remédios contra a obesidade", disse um dos autores, John Clapham, da empresa farmacêutica SmithKline Beecham, que fez o estudo em colaboração com a Universidade de Cambridge, Reino Unido.

Drogas baseadas no UCP-3 teriam pouco em comum com os moderadores de apetite usados hoje. "Elas funcionariam do outro lado da equação", disse Clapham. Em vez de reduzir a ingestão de energia, aumentariam seu consumo pelo corpo _o que, atualmente, é conseguido com o aumento de exercícios físicos.

Clapham não prevê fórmulas mágicas para os sedentários. Ele afirmou que dieta e exercícios devem continuar a protagonizar tratamentos para emagrecer. É preciso saber, agora, se há um limite para a superativação do gene _e quais os seus efeitos colaterais.

8. CIENCIA_2002_22023

A maioria dos cientistas concorda que os asteróides com mais de um quilômetro de diâmetro, capazes de destruir civilizações inteiras, são uma preocupação que só se justifica a cada punhado de dezenas de milhões de anos. Mas novos cálculos mostram que bólidos mais modestos, com 50 metros de diâmetro e a capacidade de destruir uma cidade, despencam do céu uma vez por milênio.

Na verdade, trata-se de boa notícia. Estimativas anteriores sugeriam que um evento desses ocorresse em média a cada 200 ou 300 anos. Os novos cálculos, aprimorados com o uso de informação antes mantida secreta pelo governo americano, oferecem uma estimativa mais precisa sobre a periodicidade desses episódios.

Durante os últimos oito anos, uma rede de satélites do Departamento de Defesa dos EUA tem monitorado a atmosfera terrestre com o objetivo de detectar explosões _obviamente na tentativa de monitorar o uso de armamento nuclear ao redor do globo.

Registros de bomba atômica nunca apareceram, mas, em compensação, o sistema foi capaz de apontar diversos eventos de explosões _todas causadas pela entrada de pequenos asteróides na atmosfera da Terra e sua subsequente quebra pelo atrito com o ar. Para os militares a coisa acabou não sendo lá muito útil, mas os dados se tornaram um prato cheio para os astrônomos.

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

Incidências de rochas espaciais de poucos metros de diâmetro na atmosfera acontecem com razoável frequência _anualmente, segundo os pesquisadores. "Esses corpos medidos em metros são interessantes cientificamente, mas não oferecem absolutamente nenhum perigo aos humanos", diz Robert Jedicke, da Universidade do Arizona, escolhido pela "Nature" para comentar o estudo.

A ameaça só existe quando os bólidos têm 50 metros ou mais. Foi um meteoro desse tipo (ou um disco voador, segundo fãs de ufologia) que explodiu sobre Tunguska, na Sibéria, em 1908, destruindo centenas de quilômetros quadrados de floresta. Se um desses explodisse sobre uma região habitada, poderia matar milhões. Felizmente, com base na nova estimativa, parece haver ainda nove séculos para catalogar os pedregulhos espaciais e se preparar para futuras colisões.

9. CIENCIA_2003_24219

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

A idéia foi lançada pelo ministro Roberto Amaral (Ciência e Tecnologia) e detalhada ontem durante a abertura do 1º Congresso Internacional de Biodiesel, realizado em Ribeirão Preto e promovido pela USP (Universidade de São Paulo) da cidade.

A intenção do governo é usar parte da soja transgênica já plantada no país, e que está com seu consumo proibido, na produção do combustível.

Cálculos iniciais do ministério apontam que o programa nacional do biodiesel pode representar uma economia anual de R\$ 1,8 bilhão de litros de diesel importado pelo Brasil e gerar 200 mil empregos no campo.

Francelino Grando, secretário de Política Tecnológica Empresarial do MCT (Ministério da Ciência e Tecnologia), disse que a proposta é "uma equação lógica". "Temos que ter em mente que a soja transgênica não desaparecerá no próximo ano. E precisamos ter uma alternativa."

O ministro da Agricultura, Roberto Rodrigues, afirmou que o uso da soja transgênica "é uma boa idéia". Ele defendeu até que se continue produzindo esse tipo de soja para "esmagá-la" e transformá-la em biodiesel.

Essa proposta, ainda segundo o ministro da Agricultura, será discutida pelo governo. "Assim que tivermos uma posição, cada ministério vai tratar de sua parte", afirmou Rodrigues.

O secretário do MCT também defendeu a manutenção da produção dos transgênicos. "Um assunto produtivo não pode ser tratado pela polícia", afirmou.

Francelino Grando e Roberto Rodrigues chegaram ao evento de Ribeirão num microônibus movido a biodiesel. "É para mostrar que isso é uma realidade, que não é um sonho nem um discurso", afirmou Rodrigues.

O projeto, desenvolvido pela USP de Ribeirão, consegue produzir o biodiesel a partir da mistura de óleo vegetal _incluindo o de soja_ e etanol, álcool derivado da cana-de-açúcar.

A tecnologia difere do biodiesel utilizado em outras partes do mundo, que usa o metanol _um derivado do petróleo_.

Ainda ontem, o prefeito de Ribeirão, Gilberto Maggioni (PMN), assinou uma carta de intenção com a USP para colocar parte da frota da administração municipal movida a biodiesel, já a partir de junho.

10. CIENCIA_2004_26415

Para um desavisado parece até obsessão freudiana, mas Hendrik Poinar está pedindo a todos os seus conhecidos a maior quantidade de fezes possível _quanto mais velhas, melhores_. Bioantropólogo da Universidade MacMaster, no Canadá, está prestes a investigar a relação entre neandertais e humanos modernos olhando não para seus crânios, mas para o que eles defecavam _e para as moléculas que podem contar a história deles, ocultas ali dentro_.

"Estamos recolhendo amostras de coprólitos [fezes fossilizadas] de duas cavernas em Israel com 40 mil anos, onde provavelmente Cro-Magnons [os primeiros humanos modernos] e neandertais viveram lado a lado", contou o pesquisador durante a reunião da AAAS (Associação Americana para o Avanço da Ciência).

Dadas as características muito especiais de preservação que as fezes podem alcançar, há grandes chances de elas terem preservado mais DNA do que o que se pode extrair de ossos, bem como proteínas e outras moléculas. Poinar pretende usar esse material, que segundo ele tende a ter aparência e consistência de chocolate, para extrair o máximo de informação possível sobre os dois grupos de humanos que habitaram a Palestina no fim da Era do Gelo.

Os sedimentos da caverna, que formam uma impressionante série temporal que vai até 150 mil anos atrás, também vão ser peneirados. "Depois disso, o que você faz é basicamente sequenciar tudo o que está ali e examinar toda a cadeia de relações alimentares, ecológicas e de parentesco das pessoas e animais que habitaram a caverna", afirma.

Poynar também disse estar apostando todas as suas fichas para a melhor compreensão da evolução humana na chamada paleoproteômica _o estudo das proteínas em fósseis. (RJL)

11. CIENCIA_2005_28747

Chineses e americanos enxergam o mundo de jeitos distintos _literalmente, a julgar por uma pesquisa publicada hoje. Pesquisadores da Universidade de Michigan em Ann Arbor, nos Estados Unidos, sugerem que os olhos de asiáticos tendem a ver uma imagem no seu conjunto, prestando tanto atenção ao que está em primeiro plano quanto no fundo, enquanto os americanos demoram mais o olhar no objeto central de um quadro.

"As diferenças não são minúsculas. Depois do primeiro segundo, os americanos olharam mais para o objeto central do que para o fundo durante 600 milissegundos, enquanto isso só aconteceu por 40 milissegundos com os chineses", disse à Folha Richard Nisbett, do Departamento de Psicologia da universidade.

Ele credita à sua colega Hannah Faye Chua a idéia de testar de forma visual um dado já verificado verbalmente. Pessoas nascidas na China têm mais facilidade de se lembrar de um objeto quando o vêem pela segunda vez com o mesmo fundo que aparecia na primeira olhada _o que já não acontece com os americanos.

Se isso é verdade, em que estágio da percepção ou do processamento da imagem estaria a diferença? Foi o que o grupo testou, usando óculos que rastreiam o movimento dos olhos (veja quadro à dir.). De fato, os chineses olhavam mais para o fundo, com mais intensidade e enfocando mais áreas da imagem.

Para Nisbett, diferenças culturais _principalmente na educação das crianças_ explicariam essa assimetria. "Mães americanas tendem a usar mais substantivos, e a usar mais objetos ao brincar com seus filhos pequenos. Já as chinesas e coreanas utilizam mais verbos e enfocam mais relações sociais", afirma ele. Nisbett e Chua pretendem agora ver se diferenças como essas se manifestam entre outras culturas. O estudo está na revista "PNAS".

(RJL)

12. CIENCIA_2005_28756

A boa notícia é que pesquisadores argentinos acharam o mais antigo mamífero com traços modernos da América do Sul, capaz de preencher uma lacuna de milhões de anos na história desses animais. A má é que o bicho, por enquanto, não passa de um dente.

Marcelo Tejedor, paleontólogo da Universidade Nacional da Patagônia, esboça um sorriso meio sem graça quando lhe perguntam se não é uma situação frustrante. "Estamos com um saco de 200 kg de sedimento para peneirar", diz, gesticulando para indicar o tamanho do trabalho à frente. "Quem sabe não encontramos mais alguma coisa?", afirmou durante o 2º Congresso Latino-Americano de Paleontologia de Vertebrados, encerrado no fim de semana passado, no Rio de Janeiro.

A descrição do único dente, um molar inferior, foi submetida para publicação numa revista científica, mas os dados preliminares sugerem que o animal pode tanto ter sido um placentário (como humanos, cães ou baleias) quanto um marsupial (como os cangurus, que carregam seus filhotes numa bolsa). "Achamos que é mais provável que ele seja um marsupial", diz Tejedor.

Se essa hipótese for verdadeira, os pesquisadores já sabem até em que grupo marsupial encaixar o caco: ele pertenceria ao grupo dos polidolopimorfos, comedores de frutas parecidos com gambás ou cuícas que hoje estão extintos. "As cúspides [elevações] do dente indicam essa dieta", afirma Tejedor.

A importância do achado, por mais incompleto que seja, vem da sua idade. Trata-se do mais antigo mamífero sul-americano do Paleoceno, o período geológico que marca o começo do "reinado" de seu grupo no planeta, logo depois da extinção dos dinossauros, há 65 milhões de anos. No período anterior, o Cretáceo (quando os dinos ainda eram a forma dominante de vertebrado terrestre), há diversos registros de mamíferos na América

do Sul, em especial na Argentina. Mas todos são formas muito primitivas, sem nenhuma relação direta com as espécies do grupo que estão vivas hoje.

A coisa muda de figura com o novo mamífero, ou o que sobrou dele. Ele foi achado em meio a sedimentos de origem marinha: pouco abaixo dele nas camadas de rocha estão mariscos fósseis que se extinguiram no fim do Cretáceo, enquanto lhe faziam companhia moluscos típicos do Paleoceno. "Isso significa que ele não é mais velho que 65 milhões de anos nem mais recente que 61,5 milhões de anos", resume o paleontólogo argentino.

A linhagem dos marsupiais e placentários (que são conhecidos pelo apelido comum de térios) se distingue justamente pelas cavidades especiais dos molares, que ajudam a triturar a comida com mais eficiência e estão presentes no espécime. Se for mesmo um marsupial, como os pesquisadores supõem, é possível que ele tenha vindo da América do Norte, onde membros do grupo aparecem bem antes no registro fóssil, durante o Cretáceo.