

A IMPORTÂNCIA DA CONSIDERAÇÃO DE ASPECTOS FUNCIONAIS DO TEXTO PARA A EFICIÊNCIA DE ANÁLISES RST

Juliana Thiesen FUCHS
(Universidade do Vale do Rio dos Sinos)
juliana.tfs@gmail.com

Maria Eduarda GIERING
(Universidade do Vale do Rio dos Sinos)
eduardag@unisinobr

RESUMO: Este artigo aborda a possibilidade de prever probabilisticamente como artigos jornalísticos de divulgação científica são organizados macro e microestruturalmente. Investiga-se em que medida é possível tratar da configuração prototípica microestrutural de artigos DC. Parte-se da noção de que a microestrutura de um texto depende funcionalmente de sua macroestrutura. Para a análise de exemplares do tipo textual em questão, é utilizado o modelo de relações elaborado pela *RST – Rhetorical Structure Theory*.

PALAVRAS-CHAVE: configuração prototípica; microestrutura; macroestrutura; *RST*; artigo de divulgação científica.

ABSTRACT: *This article approaches the possibility of probabilistically foresee how science journalism texts are macro and micro-structurally organized. We investigate how it is possible to study the micro-structural prototypical configuration of science journalism texts. We start from the notion that the microstructure of a text functionally depends on its macrostructure. For the analysis of exemplars of the textual type considered here, we use the relation model developed by RST - Rhetorical Structure Theory.*

KEYWORDS: *prototypical configuration; micro-structure; macro-structure; RST; science journalism text.*

0. Introdução

Neste artigo, defende-se a hipótese de que é possível prever probabilisticamente como artigos jornalísticos de divulgação científica organizam-se tanto macro quanto microestruturalmente. A noção de probabilidade adotada neste trabalho provém da concepção de texto do

lingüista Bernárdez (1995), para quem é possível prever probabilisticamente a configuração prototípica que um texto pode adotar em uma determinada situação comunicativa.

A configuração prototípica macroestrutural do tipo textual¹ artigo midiático de divulgação científica (doravante artigo DC) já foi abordada em outros trabalhos (Giering, 2007). Porém, pretende-se demonstrar em que medida é possível abordar também a configuração prototípica microestrutural de artigos DC.

Para tanto, partirei da noção de que a microestrutura de um texto depende funcionalmente de sua macroestrutura (Bernárdez, 1995; van Dijk, 1998). Para a análise de exemplares do tipo textual em questão, utilizarei o modelo elaborado pela *RST – Rhetorical Structure Theory* (Mann e Thompson, 1988).

A seguir, explicarei os conceitos e pressupostos de cada teoria abordada neste trabalho, e mostrarei a análise *RST* de um texto, procurando evidenciar como a macroestrutura do texto em questão influencia sua microestrutura e como esse fenômeno acontece em outros artigos DC.

1. A Teoria da Estrutura Retórica

A *RST* é uma teoria formulada originalmente para fins de geração automática de textos, mas que se mostrou eficiente em estudos de análise textual. A teoria procura descrever a estrutura hierárquica do texto a partir de um modelo de relações que se estabelecem recursivamente entre partes consideradas núcleos e satélites.

A coerência textual é um princípio fundamental para a *RST*, que, como teoria organizacional, atribui um papel a cada parte do texto; afinal, se cada parte não desempenhasse um papel no texto inteiro, o texto possuiria uma descontinuidade (*nonsequitur*). Assim, o objetivo da teoria é descrever a estrutura do texto, ou seja, que espécie de partes ele possui e os princípios de combinação de todas as suas partes.

Para tanto, a *RST* parte de alguns princípios. Durante o processo de elaboração da teoria, após analisar mais de 400 textos de vários tipos e tamanhos, os autores perceberam que muitos fenômenos da estrutura textual envolvem pares de partes do texto e que há relações recorrentes que acontecem entre as partes, podendo se estabelecer entre partes de diversos tamanhos, de orações a grupos de parágrafos.

¹ Neste trabalho, utiliza-se o termo *tipo textual* em vez do termo *gênero textual* seguindo-se a nomenclatura da Linguística Textual. Utiliza-se esse termo na acepção de Bernárdez (1995), para quem os tipos de texto são configurações prototípicas estáveis que os textos adotam em situações comunicativas dadas.

Além disso, perceberam que há uma assimetria nas relações entre partes do texto: em cada relação, há uma parte mais central, chamada *núcleo*, e uma mais periférica, chamada *satélite*; o núcleo e o satélite juntos formam a relação.

Além desses princípios teóricos, os autores da RST estabeleceram alguns pressupostos que embasam a teoria. Um deles é a organização, ou seja, o fato de que o texto consiste em partes funcionalmente significantes, que são combinadas para formar partes maiores e textos inteiros. Outros pressupostos são a unidade e a coerência, ou seja, o fato de que todas as partes do texto contribuem para uma única finalidade do produtor textual, isto é, são criadas para alcançar um único efeito.

Além disso, a RST pressupõe a homogeneidade: há um grupo de padrões estruturais (esquemas de relações) disponível para a organização do texto em cada escala, desde a maior até a menor. Um pressuposto fundamental da RST é a hierarquia, ou seja, o fato de que o texto é organizado de forma que as partes elementares são combinadas em partes maiores, que, por sua vez, são combinadas em partes maiores ainda até chegar à escala do texto inteiro (essa estrutura hierárquica do texto pode ser observada na figura 3 deste artigo).

Esse pressuposto contrasta com a noção de que a estrutura textual é formada por, por exemplo, correntes linearmente relacionadas de orações ou proposições semânticas.

Conforme Mann e Thompson (1988), o primeiro passo na análise de um texto é dividi-lo em unidades. O tamanho da unidade é arbitrário para a RST, podendo abranger desde itens lexicais típicos até parágrafos inteiros, ou unidades ainda maiores. Quem determina essas unidades é o analista. Afinal, as unidades previstas pela RST são relacionais, ou seja, são proposições não determinadas, e sim inferidas, que surgem da estrutura do texto no processo de sua interpretação.

O passo seguinte na análise textual é estabelecer relações entre as unidades. Para tanto, o analista deve verificar se a definição da relação plausivelmente aplica-se às unidades textuais. Um exemplo de definição de relação é o seguinte (Mann, Matthiessen e Thompson, 1992)²:

² Todas as citações em língua estrangeira presentes neste artigo foram traduzidas por esta autora.

Nome da relação: EVIDENCE

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

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

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

Efeito: a crença do leitor no núcleo é aumentada.

Locus do efeito: núcleo.

Todas as definições de relações apresentam esse mesmo conjunto de condições para o núcleo, para o satélite e para a combinação de ambos (a não ser em casos de relações multinucleares, em que as condições são determinadas para a união dos diversos núcleos). O campo *efeito* mostra o caráter funcional das relações *RST*. Isso porque, em uma análise, o relevante é a categoria do efeito que as relações produzem.

O analista pressupõe esse efeito a partir de suas considerações sobre as possíveis finalidades do produtor textual, sobre as plausíveis suposições do produtor a respeito do leitor, e sobre seu conhecimento de determinados padrões proposicionais em relação ao conteúdo do texto. "As relações da estruturação do texto refletem as opções do produtor de organização e apresentação; é nesse sentido que a *RST* é 'retórica'" (Mann, Matthiessen e Thompson, 1992: 45). O campo *locus* do efeito indica onde incide o efeito da relação.

No caso da definição exemplificada, o efeito recai apenas sobre o núcleo; isso significa que, na relação EVIDENCE, o satélite serve para aumentar alguma inclinação do leitor para o núcleo. Esse é o caso de todas as relações pragmáticas da *RST*, denominadas relações de apresentação. No caso das relações semânticas, denominadas relações de conteúdo, o *locus* do efeito está tanto no núcleo quanto no satélite, pois é necessária a compreensão de ambos para que o leitor reconheça a relação entre eles.

Na figura 1, a seguir, é possível observar a ocorrência da relação EVIDENCE:

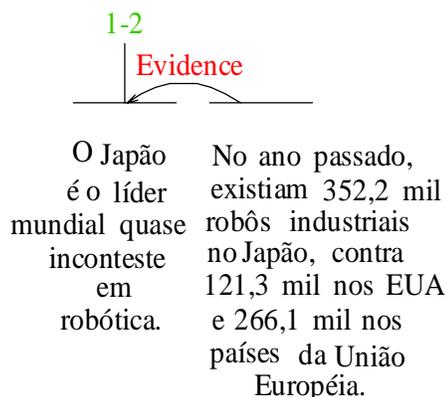


Figura 1: exemplo da relação EVIDENCE, retirado de um trecho de texto do *corpus* Summit.

Essa figura foi criada em uma ferramenta disponibilizada para representar análises RST, a *RSTTool* (O'Donnell, 2000). A ferramenta permite que o analista segmente o texto e construa sua estrutura hierárquica, sem automatizar o processo de nenhuma forma. Na figura 1, é possível observar que o conteúdo apresentado no satélite, segmento 2, pode fazer com que o leitor aumente sua crença na afirmação apresentada no núcleo, segmento 1. Assim, pode ser atribuída a esse trecho uma relação EVIDENCE.

A RST é utilizada em diversos projetos que envolvem geração e sumarização automática de textos. Um deles é o projeto ProCaCoSA (Processamento de Cadeias de Correferência para a Sumarização Automática de Textos em Português), desenvolvido pelo NILC (Núcleo Interinstitucional de Linguística Computacional) na UFSCar (Universidade Federal de São Carlos), em parceria com o PLN (Processamento de Língua Natural) do PIPCA (Programa Interdisciplinar de Pós-Graduação em Linguística Computacional) na UNISINOS (Universidade do Vale do Rio dos Sinos).

Nesse projeto, pretende-se analisar em que medida a associação de modelos – cadeias de correferência e estruturas RST – pode assegurar a melhora da coerência de sumários produzidos automaticamente.

O uso do modelo RST nesse projeto deve-se ao fato de que a estruturação de sumários que visa à produção de textos coerentes, tratando-se de um processo fundamental e baseado em conhecimento discursivo, deve ser explorada sob o aspecto não só de seleção do conteúdo, como também organizacional. A produção de sumários com a RST consiste na identificação dos núcleos e na supressão dos

respectivos satélites, para que se mantenham, no sumário gerado (um extrato do texto), apenas as informações nucleares, essenciais.

O projeto ProCaCoSA vale-se de um *corpus*, intitulado Summit, composto por 50 textos retirados do caderno de Ciência do jornal Folha de São Paulo, constituindo um sub-*corpus* do *corpus* PLN-BR Gold³. Minha participação nesse projeto consistiu na análise dos textos do *corpus* na *RSTTool*. Os primeiros 12 textos foram analisados juntamente com outro analista, que utilizou sua experiência e os textos analisados para compor sua dissertação de mestrado.

Durante o processo de análise dos 12 primeiros textos, houve discussões entre os analistas a respeito de todas as etapas da análise: a segmentação dos textos, a atribuição de relações aos segmentos e a estruturação dos textos. Dessas discussões, foram estabelecidos alguns parâmetros para a análise dos textos seguintes.

A segmentação dos textos seguiu preferencialmente os critérios estabelecidos por Carlson e Marcu (2001), para quem os segmentos devem ser basicamente orações, com algumas exceções (trechos, ainda que não-oracionais, que estão entre aspas, parênteses ou travessões, ou que aparecem sinalizados por marcadores discursivos "fortes" como "porque", "pois" etc., também são considerados segmentos).

O conjunto de relações utilizado é aquele estabelecido em Pardo (2005). Esse conjunto foi adotado por ser aquele utilizado no analisador discursivo automático DiZer (Pardo, op.cit.), para o qual foi feita uma análise de *corpus* (textos científicos do domínio da Ciência da Computação), com o objetivo de identificação das relações retóricas presentes nos textos em português, bem como de seus respectivos indicadores, ou seja, marcadores discursivos e expressões indicativas.

As 32 relações propostas desse conjunto são: *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*.

O fator que mais gerou discussão entre os anotadores foi a escolha de relações. Conforme Mann e Thompson (1988), pode haver mais de uma alternativa de análise para um texto. Afinal, o papel do analista é fundamental em todas as etapas da análise, e envolve mais

³ O projeto PLN-BR (Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil) trabalha com uma série de *corpora*, entre eles o PLN-BR Gold, disponível em <http://nilc.icmc.usp.br:8180/portal/>.

do que o conhecimento das categorias de análise (relações e processos de segmentação e de estruturação): envolve os conhecimentos de cada analista sobre o tema do texto, o contexto, os possíveis leitores e a inferência de qual foi o efeito pretendido pelo produtor. Portanto, pode haver discordância entre os analistas, e o consenso é fundamental.

Porém, a observação de cada analista e o consenso entre eles não são suficientes para que se faça uma análise eficiente de um texto. Deve haver algo no texto que indique a relação que se estabelece entre as suas partes. É possível identificar as relações entre as partes dos textos a partir de marcadores discursivos. Há relações que apresentaram marcadores em quase todas as suas ocorrências nos textos do *corpus* Summit:

- a) na relação PURPOSE, em que o satélite indica propósito ou objetivo, ocorreram marcadores como "para", "com o objetivo de", "na tentativa de";
- b) na relação CIRCUMSTANCE, em que o satélite apresenta um quadro para o núcleo, houve marcadores como "quando", "antes que", "enquanto", "onde", "assim que";
- c) na relação CONCESSION, em que o núcleo apresenta uma quebra de expectativa em relação ao que é dito no satélite, houve marcadores como "mas", "apesar de", "embora", "mesmo";
- d) na relação CONDITION, em que o satélite apresenta um fato que condiciona o núcleo, ocorreram os marcadores "se", "contanto que", "caso".

Porém, esses marcadores, em geral, indicam apenas as relações que ocorrem no nível elementar dos textos (relações entre orações). As relações que ocorrem nos níveis hierárquicos superiores em geral não se mostraram indicadas por marcadores.

Assim, é possível dizer que as pistas (marcadores discursivos) da superfície dos textos não são eficientes para indicar que relações ocorrem na organização textual. Portanto, a questão permaneceu: como estabelecer um critério para atribuir relações às partes do texto em todos os seus níveis hierárquicos?

2. Microestrutura e macroestrutura: dependência funcional

Uma resposta para essa questão pode estar no uso da *RST* feito pelos projetos ORTO (Organização Retórica de Textos de Opinião)⁴ e

⁴ Finalizado. Desenvolvido no período entre janeiro de 2003 a dezembro de 2005.

ORTDC (Organização Retórica de Textos de Divulgação Científica)⁵, desenvolvidos pela Prof^a Dr^a Maria Eduarda Giering no PPGLA (Programa de Pós-Graduação em Lingüística Aplicada) da UNISINOS.

Nesses projetos, dos quais participei como bolsista de iniciação científica, utiliza-se o modelo *RST* conforme sugestão do lingüista textual Bernárdez (1995), que concebe o texto como sistema complexo, aberto e dinâmico, e considera a coerência como algo que é produzido no transcurso do processo de produção ao processo de recepção do texto. Desse modo, o texto não é um produto, não é algo estático, e sim uma ação, parte de um processo. Uma das conseqüências da complexidade do texto é o não-determinismo:

(...) não é possível prever que forma concreta adotará um texto em uma situação concreta, por mais profundo que seja nosso conhecimento das condições em que ele será produzido; podemos, porém, prever com considerável precisão a estrutura das orações, que podemos considerar "linguagem em condições de laboratório". (Bernárdez, 1995: 58).

Dessa maneira, a composição textual está submetida a um grau de indeterminação que não aparece no nível oracional ou nos níveis sub-oracionais.

O texto, sendo um sistema complexo, é, portanto, qualitativamente diferente de sistemas simples, como a oração e os sistemas sub-oracionais. Ele apresenta características como a abertura (interação com o contexto), a dinamicidade (caráter processual), a criatividade (composição de textos concretos adequados a situações concretas), a naturalidade (frente à artificialidade dos objetos simples, que são idealizados para serem estudados cientificamente), a falta de estabilidade ou falta de equilíbrio (por causa da interação com o contexto) e a dependência das condições iniciais de produção (dependência do contexto). Por isso, Bernárdez (1995) argumenta que o texto não pode ser estudado com os mesmos modelos que servem para o estudo da oração.

Segundo Bernárdez (op.cit.), a primeira palavra de uma oração abre várias vias de continuidade, ou seja, pode ser seguida por determinadas palavras, mas não por outras. Seria possível, a partir de uma palavra ou um sintagma, prever que palavra ou sintagma pode (ou não) se seguir (por exemplo, não há a possibilidade de um artigo ser seguido de um verbo ou de outro artigo). Então, conforme o autor, cada

⁵ Em andamento. Iniciado em janeiro de 2006 e com previsão de término em dezembro de 2007.

palavra abre possíveis vias, ao mesmo tempo em que restringe as possibilidades.

Entretanto, a partir de uma oração, não é possível prever exatamente que outra oração se seguirá (ou não) no texto, justamente por causa da complexidade que é instaurada a partir do nível textual: o texto não é um agrupamento de orações. Para Bernárdez (op.cit.), ao traspasar os limites da oração, deparamo-nos com um campo de indeterminação inexistente dentro daqueles e, ao mesmo tempo, com um número muitíssimo maior de vias de continuidade possível.

A indeterminação, ou seja, a imensa quantidade de possíveis continuações e a impossibilidade de valorar exatamente a probabilidade de cada uma delas, faz com que não pareça possível estabelecer sequer o número de vias que podem seguir uma oração determinada. Existe, portanto, "*um salto qualitativo radical da oração ao texto*: dentro dos limites da oração, um elemento vem determinado pelo anterior (e pelo contexto), mas esse determinismo não é em absoluto claro no texto" (Bernárdez, op.cit. 81).

Esse salto qualitativo radical entre oração e texto intensifica-se com o fato de não haver categorias universalmente aceitas para o estudo do texto. É por isso que Bernárdez (1995) sugere, para o estudo do texto, o modelo desenvolvido pela *RST – Rhetorical Structure Theory* (Mann; Thompson, 1988), teoria explicada anteriormente.

Assim, o autor propõe a suposição de que "a organização textual pode ser entendida como uma série de vias ou opções de continuidade, etiquetadas aqui com as relações apresentadas pela *RST*" (Bernárdez, op.cit: 85).

Contudo, mesmo com a possibilidade de um modelo para o estudo do texto, continua havendo uma transição "catastrófica" na passagem do nível da oração para o nível do texto. É por isso que, segundo o mesmo autor, para estudar o texto, deve-se estabelecer uma nova teoria da linguagem, o que é possível a partir de estudos como a *Teoria das Catástrofes*.

A catástrofe é uma mudança qualitativa no sistema: é a passagem de um estado de equilíbrio a outro, determinada pela atuação de variáveis de controle, geralmente externas; esses estados de equilíbrio seriam definidos por um protótipo, um atrator, que define um estado estruturalmente estável, ou estado ótimo, a que um sistema almeja chegar, no contínuo do processo.

No caso do texto, a estabilidade estrutural aparece, de acordo com Bernárdez (1995), quando em situações semelhantes se produzem textos semelhantes. Essas invariâncias podem ser chamadas de tipos de

texto, ou seja, formas mais ou menos prototípicas que os textos adotam em situações comunicativas dadas.

Os tipos de texto são estados de equilíbrio determinados por um conjunto de variáveis de controle; um tipo de texto transforma-se em outro quando há uma modificação nas variáveis de controle, como o contexto.

Na visão de Bernárdez (op.cit:134), não existe nunca uma única forma de produzir um texto, mesmo que as condições contextuais sejam praticamente idênticas; além disso, é impossível haver dois contextos idênticos. Assim, "nunca poderemos chegar a explicar, a não ser de maneira relativamente vaga e imprecisa, a formação de um texto concreto (e sua recepção ou compreensão)".

Essa dependência dos fatores externos do contexto é, o que as ciências do caos chamam de sensibilidade às condições iniciais, e que suscita a seguinte questão: como o produtor pode estar seguro de que seu texto garante, na medida do possível, a compreensão do interlocutor? A coerência pode ser entendida em termos de forma ótima que deverá adotar o texto para solucionar esse problema (Bernárdez, op.cit.).

O autor referido procura explicar os processos necessários para se alcançar o "texto ótimo". Um desses processos é a retroalimentação, que serve para orientar o processo de comunicação lingüística. É impossível garantir o sucesso de uma comunicação lingüística; o máximo que os participantes da comunicação podem fazer é acercar-se o máximo possível ao ideal de texto, ao estado estável, ao estado ótimo, ao protótipo.

No caso da comunicação escrita, não há possibilidade de retroalimentação continuada, ou seja, não há meios de confirmar que o texto que se está produzindo é o ótimo. Nesse caso, o produtor deve recorrer a todo o seu conhecimento, inclusive o que pode saber de antemão sobre os possíveis contextos de recepção e sobre as características imagináveis do interlocutor (leitor tipo), para que seu texto conte com as maiores garantias possíveis: o produtor deve buscar um leitor e um contexto prototípicos, ou seja, o mais estável possível.

Na visão de Bernárdez (1995), são esses textos, os que não podem contar com a retroalimentação do interlocutor, que estão mais bem construídos, formados com maior cuidado. Ademais, um texto (de determinado tipo), quanto mais cuidadoso for, mais será previsível, pois mais coincidirá com o texto prototípico. Assim, a configuração prototípica de um (tipo de) texto é, em consequência, a configuração ótima, e, portanto, a mais coerente.

É por isso que alguns tipos de texto são mais automáticos que outros; ou seja, estão mais fixados culturalmente, permitem menos variação individual, apresentam uma configuração prototípica muito estável, dotada de uma probabilidade muito alta (por exemplo, cartas comerciais ou receitas culinárias).

A causa dessa diferença entre os tipos de texto é a maior ou menor previsibilidade dos fatores contextuais (previsibilidade essa que permite ao produtor construir seu texto em função do protótipo, do exemplar ótimo): a informação do produtor sobre o leitor de uma receita culinária é muito alta (muito mais que no caso de um artigo jornalístico de opinião, por exemplo), e a sua busca por um protótipo é simples, quase automatizada.

Para Bernárdez (op.cit.), se as diversas variáveis que estão em jogo na produção de um texto – contexto, interlocutor – forem conhecidas, é possível prever probabilisticamente como esse texto vai se organizar. Afinal, é possível prever as estratégias mais prováveis que o produtor adotará para que seu texto seja “ótimo”, prototípico, o mais estável possível.

Porém, a predição sobre a coerência do texto – ou seja, sobre a forma ótima que o texto adotará em uma determinada situação comunicativa – tem um caráter qualitativo. Isso porque a forma ótima que o texto vai adotar dependerá de fatores contextuais diferentes de um caso a outro.

O máximo a fazer é estabelecer uma relação entre os diversos fatores envolvidos na produção e na compreensão dos textos e observar sua tendência geral de desenvolvimento em função da variação desses fatores:

(...) não é possível dizer que um texto será coerente se reúne as condições *a, b, c, ... n* de maneira absoluta e em forma independente do contexto concreto no qual foi produzido. Podemos estabelecer, em forma também *qualitativa*, as relações entre os subsistemas de forma que o aumento das variáveis de um leve consigo uma variação no outro. (...) Isso obriga a utilizar para o estudo do texto ferramentas metodológicas diferentes das habituais em lingüística; o estudo *empírico* (sobre textos reais produzidos/recebidos em condições reais etc.) resulta agora fundamental como via para a generalização cientificamente válida (Bernárdez, 1995: 146).

Esse estudo empírico pode envolver a noção de tipo de texto. Conforme dito anteriormente, o tipo de texto é uma configuração prototípica que um texto adota em uma determinada situação

comunicativa: “a configuração prototípica pode ser considerada como aquela configuração que tem mais probabilidades, de acordo com a experiência pessoal e interpessoal, de conseguir o resultado desejado no entorno específico” (Bernárdez, op.cit: 188). Portanto, a configuração prototípica do texto depende da configuração prototípica do contexto.

Foi esse tipo de estudo empírico que os projetos ORTO e ORTDC procuraram fazer: o objetivo dos projetos foi verificar a configuração prototípica macroestrutural, a partir das relações da *RST*, de artigos jornalísticos de opinião e artigos DC midiáticos. Dessa forma, os projetos puderam verificar, em *corpora* (150 artigos de opinião e 120 artigos DC), a recorrência das relações *RST* entre as macroproposições dos textos.

A partir dessa verificação, é possível uma previsão probabilística de como artigos jornalísticos de opinião e artigos DC midiáticos podem organizar-se macroestruturalmente. O uso do modelo *RST* conforme a teoria de Bernárdez (1995) é possível porque, segundo o autor, a noção de “retórica” da *RST* tem a ver com o que ele chama de “macroestruturação textual”. Afinal, para Bernárdez (1995), como o texto é uma ação, seu processamento é estratégico.

Então, o autor, ao procurar estudar não apenas as estruturas textuais, mas os processos que subjazem a essas estruturas, explica o processo de produção textual dizendo que o produtor tem à sua disposição várias opções para macroestruturar seu texto e opta por cada uma delas estrategicamente, em função de seu objetivo e do contexto.

Bernárdez (1990) comenta que a noção de “retórica” da *RST* está intimamente ligada a essa noção de “macroestruturação textual”. Dessa forma, cada relação proposta pela *RST* seria uma estratégia à disposição do produtor para macroestruturar seu texto. Esse processo de macroestruturação, segundo Bernárdez (op.cit.), afeta primeiro as macroproposições do texto, depois as macroproposições de nível inferior, até afetar o nível mais baixo do texto, representado pelas proposições semânticas.

A macroestruturação do texto, comentada em Bernárdez (op.cit.), remete às noções de macroestrutura e macroproposição de van Dijk (1998). Para esse autor, a macroestrutura é uma estrutura especial de tipo global, de natureza semântica, que constitui a representação abstrata da estrutura global de significado de um texto.

Essa representação abstrata tem a ver com a capacidade do falante de atribuir um tema ou assunto a um texto depois da leitura, e constitui um significado global, que se dá não em termos de orações individuais, mas sim de seqüências completas de orações. Afinal, de

acordo com o mesmo autor, só as seqüências de orações que possuem uma macroestrutura podem ser denominadas textos.

A macroestrutura relaciona-se com a microestrutura, ou seja, a estrutura local de um discurso ou texto: a estrutura das orações e as relações de conexão e de coerência entre elas. Portanto, para que um texto seja considerado como tal, deve obedecer não só às regras de coerência local, mas também às de coerência global, que se observam na macroestrutura textual.

Segundo van Dijk (1998), da mesma forma que a microestrutura é formada por proposições, a macroestrutura é formada por macroproposições. Cada macroproposição seria um conjunto de proposições que têm uma unidade semântica e que podem, a partir de um processo cognitivo de redução da informação semântica, ser reduzidas a uma única proposição.

Esse processo de redução da informação semântica é feito por meio de *macrorregras*, ou seja, regras (como supressão, generalização, construção etc.) que vinculam as proposições da microestrutura textual às da macroestrutura textual.

Cada macroproposição é relativa às proposições do nível imediatamente inferior; desse modo, o texto está formado por vários níveis macroproposicionais, sendo o nível mais alto composto de uma única macroproposição que dá conta do tema do texto inteiro, conforme a figura 2:

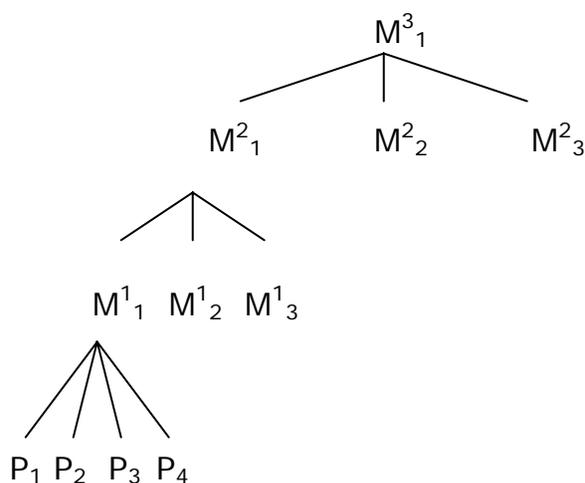


Figura 2: Representação esquemática da macroestrutura semântica (van Dijk, 1996, p. 57).

Essa árvore ilustra a relação entre as proposições e as macroproposições em diversos níveis. O nível inferior, composto pelas

proposições (P1, P2, P3, P4), apresenta todas as informações do texto, e cada nível superior apresenta macroproposições com redução da informação. O nível superior, composto pela macroproposição geral do texto (M^3_1), apresenta uma única informação que dá conta do tema ou assunto principal do texto.

O processo cognitivo de formação das macroproposições depende das funções pragmáticas do texto, denominadas *macroatos de fala*: da mesma forma como os atos de fala regulam as condições de adequação das seqüências de proposições, há atos de fala mais globais, os macroatos de fala, que orientam as seqüências de macroproposições. Dessa forma, pode ser atribuído um macroato de fala à macroproposição geral do texto, isto é, é possível dizer que o texto inteiro foi produzido a partir de uma única finalidade.

É por isso que, de acordo com van Dijk (1998), a coerência local depende da coerência global do texto, e só é possível compreender as proposições do texto a partir da sua funcionalidade em relação às macroproposições. Isso tem a ver com o que Bernárdez (1995) chama de níveis estratégicos: o texto seria uma ação global que se desdobra em ações menores, até chegar aos níveis menos estratégicos e mais automatizados (os níveis sub-oracionais, por exemplo).

A formação de macroproposições a partir das macrorregras também depende do tipo de texto em questão. Afinal, em um determinado texto, alguma informação pode ser mais importante ou pertinente que a mesma informação em outro texto. Por exemplo, considere-se a informação da descoberta de que as diferentes categorias raciais da humanidade são construções sociais e não têm qualquer respaldo genético.

Essa informação pode ser nuclear em um artigo de divulgação científica, que, em geral, tem como finalidade divulgar um novo conhecimento científico. A mesma informação, porém, pode ser satélite em um artigo de opinião, que, em geral, tem como finalidade defender um ponto de vista; nesse caso, a informação poderia constituir um argumento do produtor textual.

Portanto, a aplicação cognitiva das macrorregras para a formação das macroproposições pode depender da *superestrutura* textual: a estrutura esquemática, global, abstrata de um texto, que é reconhecida cognitivamente pelo leitor, e que define a ordenação global do texto e as relações hierárquicas de seus respectivos fragmentos.

Nos projetos ORTO e ORTDC, optou-se por analisar apenas o nível mais alto de organização do texto, o macroestrutural, sem a consideração dos níveis hierárquicos inferiores. Afinal, procurou-se investigar a organização retórica global do texto. Dessa forma, cada

relação *RST* corresponderia a uma das macroações (Bernárdez, 1990) do produtor para organizar seu texto globalmente, em dependência do tipo textual e do macroato de fala. Portanto, as unidades de análise dos projetos foram segmentos contíguos de texto reduzíveis a uma macroproposição.

Nesses projetos, observou-se que os textos apresentam uma organização macroestrutural prototípica: no projeto ORTO, verificou-se que a seqüência de relações macroestruturais mais freqüente nos artigos de opinião é EVIDENCE-EVALUATION. No projeto ORTDC, verificou-se, nos artigos DC analisados, a recorrência da seqüência de relações PREPARATION-SUMMARY-ELABORATION-INTERPRETATION-COMMENT.

Com essas constatações, os projetos ORTO e ORTDC mostraram a possibilidade de se prever probabilisticamente a organização global dos tipos textuais contemplados.

Análises *RST* no nível macroestrutural são possíveis porque, conforme dito anteriormente, a teoria prevê que as relações se estabeleçam entre proposições não determinadas, e sim inferidas, que surgem da estrutura do texto no processo de interpretação de textos.

A unidade elementar de análise adotada nos projetos ORTO e ORTDC, por ser inferida, não está determinada diretamente no texto, mas é definida somente após o processo de transformação das proposições em macroproposições por meio das macrorregras.

Isto é, a unidade elementar de análise adotada nesses projetos depende de um processo de inferência realizado pelo analista após a leitura integral do texto. É por isso que, nos projetos ORTO e ORTDC, a análise *RST* representa a organização macroestrutural dos textos.

Durante a participação no projeto ProCaCoSA, os analistas perceberam que, além da organização macroestrutural do textos, as análises *RST* deveriam representar a sua organização microestrutural: como as macroproposições dos vários níveis e as proposições relacionam-se entre si, formando uma estrutura hierárquica.

Logo, em vez de apenas seguir os princípios metodológicos previstos para análises *RST* - segmentação do texto em orações e atribuição de relações a elas para compor segmentos progressivamente maiores -, os analistas deveriam tratar todos os níveis hierárquicos do texto da mesma forma como foi abordado o nível macroestrutural nos projetos ORTO e ORTDC.

Da mesma forma que cada macroproposição corresponde a uma macroação do produtor textual para organizar seu texto globalmente, as macroproposições de nível inferior e as proposições correspondem a ações menores do produtor. Desse modo, em cada nível hierárquico, as

relações *RST* representam alguma ação que contribui para a organização textual.

Porém, as ações não são desempenhadas separadamente em cada nível hierárquico do texto. Na verdade, as ações dos níveis menores são condicionadas pelas dos níveis maiores. Afinal, Bernárdez (1995) considera o texto como uma unidade formada por outras unidades menores, da mesma forma que uma ação complexa está dividida em (sub) ações parciais.

Conforme o autor, em cada nível, são utilizadas certas estratégias para conseguir os objetivos parciais correspondentes. Além disso, conforme dito anteriormente, para van Dijk (1998), a coerência local depende da coerência global do texto, e só é possível compreender as proposições do texto a partir da sua funcionalidade em relação às macroproposições.

A partir dessas considerações, este trabalho partiu dos seguintes questionamentos: se é possível dizer que as partes do texto têm funcionalidade em relação ao todo, então como se dá essa relação entre o todo e as partes em análises *RST*? De que forma as relações da microestrutura textual dependem das relações da macroestrutura? Além disso, já que é possível prever probabilisticamente como se configuram as relações *RST* entre as macroproposições de artigos DC, seria possível prever probabilisticamente também a configuração das relações entre as macroproposições de outros níveis hierárquicos e as proposições?

3. Artigo DC: organização macro e microestrutural

Para responder a esses questionamentos, foram considerados apenas os 12 primeiros textos do *corpus* Summit analisados na *RSTTool*. Nesses textos, foram observadas recorrências de relações *RST* tanto no nível macroestrutural quanto nos níveis hierárquicos inferiores. A figura 3 mostra a análise *RST* de um dos textos do *corpus* Summit, que exemplifica as principais recorrências observadas:

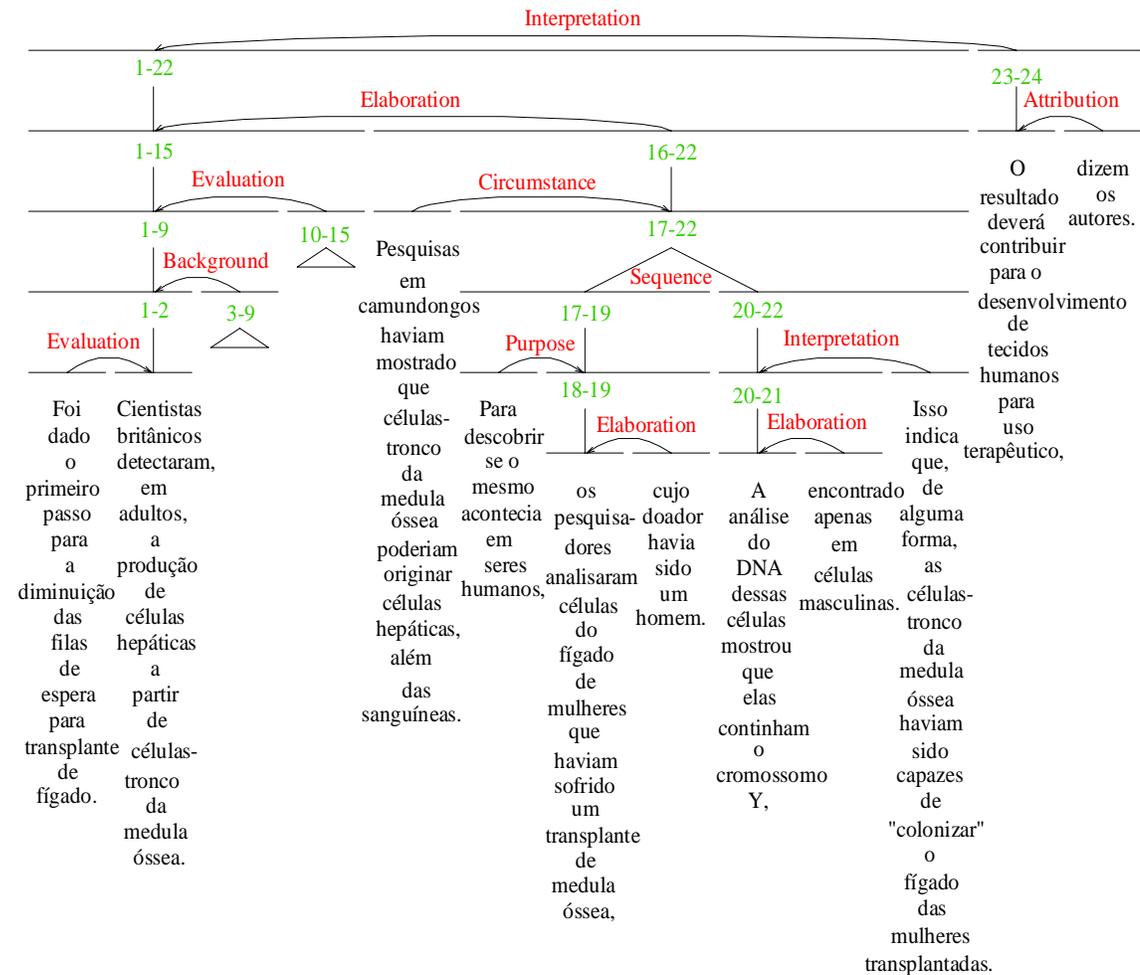


Figura 3: Estrutura hierárquica de texto do *corpus* Summit representada na *RSTTool*.

Entre as recorrências de relações *RST* observadas nos 12 textos do *corpus* Summit, serão comentadas primeiramente as concernentes ao nível macroestrutural dos textos.

A incidência de determinadas relações *RST* nos níveis macroestruturais dos textos analisados está ligada à superestrutura do tipo textual em questão (artigo DC midiático). O macroato de fala desse tipo textual é divulgar uma pesquisa a um público variado. Portanto, os textos desse tipo apresentam determinados elementos em sua superestrutura:

- a) a menção à pesquisa divulgada;
- b) a apresentação dos procedimentos metodológicos utilizados na pesquisa divulgada;

- c) a avaliação e a interpretação dos pesquisadores ou de outros membros da comunidade científica sobre a repercussão da pesquisa.

Esses elementos superestruturais foram observados nos 12 primeiros textos do *corpus* Summit, e, para cada um deles, os analistas convencionaram atribuir determinadas relações.

A menção à pesquisa divulgada é sempre o segmento nuclear mais alto na estrutura hierárquica dos textos, já que a finalidade do tipo textual em questão é divulgar alguma pesquisa. Todas as relações macroestruturais de cada texto estão ligadas como satélites a esse segmento nuclear⁶. É o que pode ser visto na figura 3, entre os segmentos 1-15, núcleo macroestrutural do texto, e os segmentos 16-22 e 23-24, satélites de relações macroestruturais.

Provavelmente isso se deve ao fato de que, por constituírem artigos DC jornalísticos, esses textos apresentam, assim como na grande maioria das notícias e reportagens, a informação principal já no início do texto, desenvolvendo essa informação ao longo do texto. É o que Massarani e Moreira (2005) chamam de texto "piramidal", em que a novidade da pesquisa é o ponto de partida do autor da matéria divulgativa.

A apresentação dos procedimentos metodológicos utilizados na pesquisa divulgada (ou os procedimentos observados nos objetos pesquisados) geralmente constitui um segmento macroestrutural diretamente ligado ao segmento nuclear mais alto 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 para referir procedimentos de pesquisa em um nível macroestrutural é apresentado na figura 3, como pode ser observado entre os segmentos 1-15 e 16-22.

A interpretação e a avaliação dos pesquisadores ou de outros membros da comunidade científica sobre a repercussão da pesquisa geralmente constitui um segmento macroestrutural diretamente ligado ao segmento nuclear mais alto na estrutura hierárquica do texto.

As relações observadas nesse tipo de segmento são INTERPRETATION (em 33% dos textos) e EVALUATION (em 25% dos textos). Um exemplo de ocorrência da relação INTERPRETATION para referir macroestruturalmente a posição de pesquisadores sobre a

⁶ Nos diagramas arbóreos representados na *RSTTool*, os núcleos e os satélites são indicados por setas, que apontam sempre do satélite para o núcleo, exceto no caso de relações multinucleares.

repercussão da pesquisa divulgada pode ser observado na figura 3, entre os segmentos 1-22 e 23-24.

Além dessas recorrências no nível macroestrutural dos textos, os analistas verificaram recorrências de relações *RST* também nos níveis hierárquicos inferiores: as relações *RST* que ocorrem no nível macroestrutural dos textos organizam-se, internamente, por meio de seqüências recorrentes de determinadas relações. Da mesma forma, verificou-se que algumas relações que ocorrem em níveis hierárquicos inferiores freqüentemente aparecem subordinadas a determinadas relações macroestruturais.

Afinal, as características superestruturais do tipo de texto em questão restringem as possibilidades de organização interna das relações *RST*: a base da pirâmide, composta das informações principais do texto (no caso de artigos DC, a divulgação de uma pesquisa e/ou de uma descoberta científica), organiza-se com relações, em geral avaliativas ou contextualizadoras, que giram em torno do núcleo principal.

O afunilamento da pirâmide, por sua vez, composto pelo detalhamento da informação nuclear (no caso de artigos DC, métodos, resultados e interpretações da pesquisa e/ou descoberta divulgada), organiza-se com várias relações que indicam processo e interpretação. Essa organização será discutida a seguir, considerando-se cada segmento macroestrutural da análise apresentada na figura 3.

Os analistas observaram que, em geral, o núcleo macroestrutural dos textos aparece composto de alguma relação que tem a função de captar o leitor para a informação que será apresentada. É o caso, na análise apresentada na figura 3, da relação EVALUATION entre os segmentos 1 e 2. O segmento 2 apresenta o fato científico divulgado no texto, e o segmento 1 apresenta um comentário avaliativo que tem a função de captar o leitor para a leitura do artigo.

Conforme dito anteriormente, nos artigos DC, o detalhamento do fato científico divulgado costuma ser apresentado pelo satélite de uma relação ELABORATION (no caso da análise apresentada na figura 3, segmentos 16 a 22).

Os analistas observaram que esse detalhamento, em geral, é composto pelas relações SEQUENCE ou LIST, indicando uma seqüência ou uma lista de procedimentos desenvolvidos pelos cientistas (no caso da análise apresentada na figura 3, o detalhamento ocorreu por meio de uma relação SEQUENCE entre os segmentos 17 a 22).

Outra recorrência macroestrutural dos artigos DC analisados, comentada anteriormente, é o fato de os textos apresentarem, após o detalhamento do fato científico divulgado, a interpretação desse fato,

concretizada pelo satélite de uma relação INTERPRETATION (no caso da análise apresentada na figura 3, segmentos 23 e 24).

Os analistas observaram que essa interpretação muitas vezes foi composta da fala de um cientista, conforme indicado pela relação CONTRIBUTION, entre os segmentos 23 e 24 na figura 3.

Essa superestrutura recorrente dos artigos DC jornalísticos mostra uma restrição das possibilidades de relações RST em cada nível hierárquico. É recorrente:

- a) que o núcleo macroestrutural dos textos seja composto pelas relações EVALUATION, INTERPRETATION ou BACKGROUND precedendo o núcleo;
- b) que a relação ELABORATION macroestrutural seja composta pelas relações LIST e SEQUENCE;
- c) que haja uma relação INTERPRETATION após a relação ELABORATION macroestrutural, e que essa relação seja composta pela relação CONTRIBUTION.

Assim, a partir da análise RST exemplificada e dos dados computados das análises de 12 textos do *corpus* Summit, é possível dizer que a macroestrutura (representada pelas relações RST nos níveis hierárquicos superiores) determina, em parte, como se dá a microestrutura (representada por relações RST em níveis hierárquicos inferiores).

Portanto, se é possível prever probabilisticamente a organização macroestrutural de artigos DC, como foi feito no projeto ORTDC, também é possível prever probabilisticamente a sua organização microestrutural.

REFERÊNCIAS BIBLIOGRÁFICAS

BERNÁRDEZ, E. *Teoría y epistemología del texto*. Madrid: Cátedra, 1995.

_____. Las macroestructuras textuales como objeto del estudio lingüístico. In: JORNADAS DE LENGUA Y LITERATURA INGLESA E NORTEAMERICANA, 1, 1989, Logroño. *Actas de las Jornadas de Lengua y Literatura Inglesa y Norteamericana*. La Rioja: Universidad de Zaragoza, 1990: 107-119.

CARLSON, L.; MARCU, D. *Discourse tagging reference manual*. ISI Technical Report ISI-TR-545, 2001.

GIERING, M. E. Configuração prototípica de artigos de divulgação científica e o texto como sistema. *Anais - IV Simpósio Internacional de*

Estudos de Gêneros Textuais. Tubarão: Universidade do Sul de Santa Catarina, 2007. v. 1. p. 1-12.

MANN, W.C.; THOMPSON, S.A. *Rhetorical Structure Theory: toward a functional theory of text organization*. *Text*, 8 (3). 1988: 243-281.

MANN, W.C.; MATTHIESSEN, C.M.I.M.; THOMPSON, S.A. *Rhetorical Structure Theory and Text Analysis*. In: MANN, W.C.; THOMPSON, S.A. *Discourse description: diverse linguistic analyses of a fund-raising text*. Amsterdam: John Benjamins, 1992: 39-78.

MASSARANI, L.; MOREIRA, I. C. *A retórica e a ciência*. Dos artigos originais à divulgação científica. *Multiciência*. Revista Interdisciplinar dos Centros e Núcleos da UNICAMP. Atualizado em 25 de maio de 2005. Disponível em: <http://www.multiciencia.unicamp.br/intro_04.htm> Acesso em: 4 de outubro de 2006.

O'DONNELL, M. *RSTTool 2.4: A Markup Tool for Rhetorical Structure Theory*. Proceedings of the International Natural Language Generation Conference (INLG'2000), 13-16 Jun. 2000, Mitzpe Ramon, Israel. 253-256.

PARDO, T. A. S. *Métodos para Análise Discursiva Automática*. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2005

VAN DIJK, T. A. *La noticia como discurso: comprensión, estructura y producción de la información*. Barcelona: Paidós, 1996.

_____. *Estructuras y funciones del discurso*. 12. ed. Madrid: Siglo Veintiuno Editores. 1998.

Recebido em setembro de 2007

Aprovado em fevereiro 2008