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

A Construção de um Corpus de Textos Científicos em Português do Brasil e sua Marcação Retórica

Thiago Alexandre Salgueiro Pardo
Maria das Graças Volpe Nunes

NILC-TR-03-08

Setembro 2003

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

Resumo

Este relatório relata a construção de um corpus de textos científicos em português do Brasil, no domínio da Computação, e sua marcação retórica segundo a *Rhetorical Structure Theory* – RST (Mann and Thompson, 1987). A construção e a marcação do corpus fazem parte do desenvolvimento de um analisador discursivo automático para o português, chamado DiZer-PBr – *DIscourse analyZER for BRazilian Portuguese*, no âmbito de um projeto de doutorado.

Este trabalho conta com o apoio financeiro da FAPESP e do CNPq.

ÍNDICE

1. INTRODUÇÃO **2**

2. CORPUSTCC **2**

3. *RHETORICAL STRUCTURE THEORY* – RST **4**

4. MARCAÇÃO RETÓRICA DO CORPUSTCC **8**

- **4.1. RST ANNOTATION TOOL** **8**
- **4.2. SEGMENTAÇÃO TEXTUAL** **9**
- **4.3. ELENCO DE RELAÇÕES RETÓRICAS** **10**
- **4.4. ESTRATÉGIA DE MARCAÇÃO RETÓRICA** **21**
- **4.5. DADOS DO CORPUS MARCADO** **21**

5. CONSIDERAÇÕES FINAIS **23**

REFERÊNCIAS BIBLIOGRÁFICAS **23**

1. Introdução

Neste relatório, são relatadas a construção de um corpus de textos científicos em português do Brasil, no domínio da Computação, e a metodologia adotada para sua marcação retórica nos moldes da *Rhetorical Structure Theory* – RST (Mann and Thompson, 1987). O corpus foi nomeado de CorpusTCC – corpus de Textos Científicos da Computação.

A construção e a marcação retórica deste corpus fazem parte do desenvolvimento de um analisador discursivo automático para o português, chamado DiZer-PBr – *DIscourse analyZER for BRazilian Portuguese*[1]. O objetivo deste sistema é, a partir de um texto, gerar suas estruturas discursivas, mais especificamente, as estruturas retórica, semântica e intencional. No nível retórico, adota-se a RST, que é uma das teorias retóricas mais difundidas atualmente; nos níveis semântico e intencional, consideram-se, respectivamente, as relações semânticas de Jordan (1992) e as relações intencionais da *Grosz and Sidner Discourse Theory* – GSDT (Grosz and Sidner, 1986). A análise do corpus marcado (segundo essas teorias discursivas) deverá fornecer o conhecimento necessário para a automação da análise discursiva proposta, ou seja, para a viabilização do DiZer.

Diante dos objetivos almejados, a marcação retórica do corpus apresentada neste relatório é, na realidade, apenas uma etapa da marcação necessária para o desenvolvimento do analisador discursivo por completo. Futuramente, o corpus deverá ainda ser marcado semanticamente e intencionalmente. Para mais detalhes sobre o desenvolvimento do DiZer, vide Pardo et al. (2003).

A próxima seção apresenta as principais características do corpus. A seguir, na Seção 3, uma breve introdução a RST é feita, para então, na Seção 4, discutir-se a marcação retórica do corpus.

2. CorpusTCC

A compilação do corpus foi guiada pelas exigências impostas para o desenvolvimento do DiZer. Dessa forma, pelo fato do DiZer estar sendo desenvolvido para análise discursiva de textos científicos da Computação, o corpus, que é a base para sua viabilização, deve ser representativo desse tipo de discurso.

Para compor o corpus, foram coletadas dos corpora do NILC (Pinheiro e Aluísio, 2003; Feltrim et. al., 2001) 47 introduções de dissertações de Mestrado e 3 introduções de qualificações de Mestrado, ambos os casos do domínio da Ciência da Computação, resultando, no total, em 50 introduções, em formato txt, com tamanho de 1 a 4 páginas. Esse corpus contém, aproximadamente, 53.000 palavras e 1.350 sentenças.

Coletaram-se somente as introduções dos textos, em vez dos textos completos, pelas seguintes razões: (a) as introduções são suficientes para cumprir a finalidade a qual o corpus serve; (b) marcar retoricamente dissertações e qualificações completas (como se discutirá na Seção 4) é inviável devido à complexidade desta tarefa e o tempo que consome.

As introduções são de várias áreas da Computação, a saber: banco de dados (bd), engenharia de software (es), hipermídia (h), inteligência computacional (ic) e sistemas distribuídos (sd). Para compor o corpus, essas introduções foram

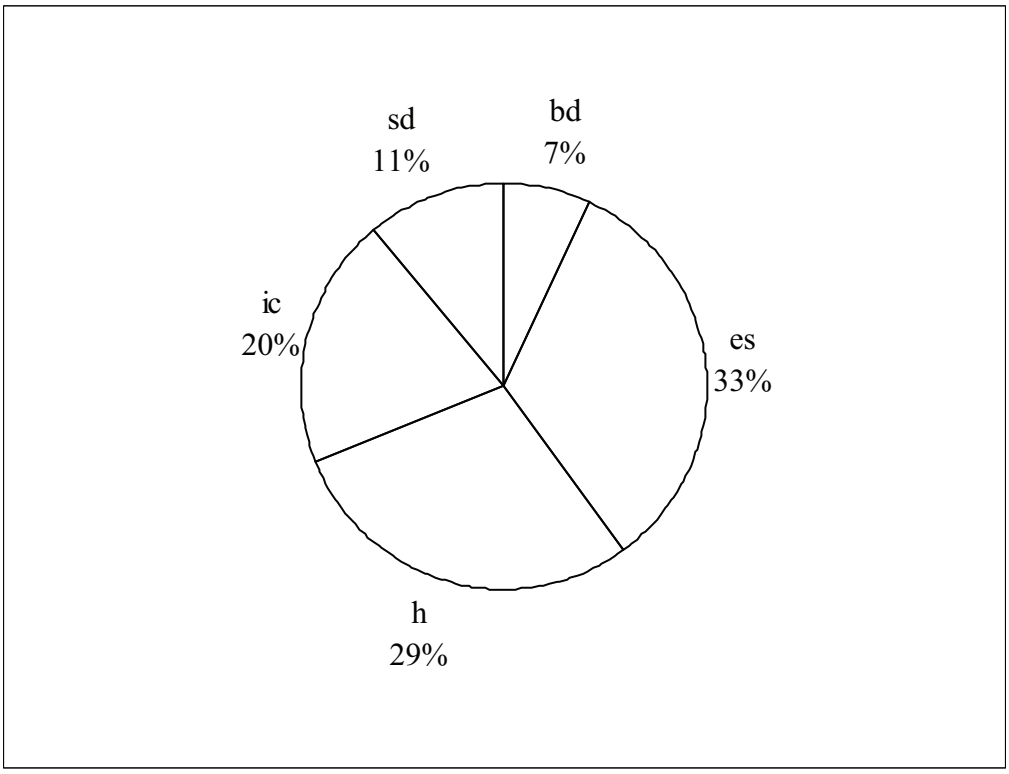
[1] http://www.nilc.icmc.usp.br/~thiago/DiZer.html

subdivididas de acordo com suas seções, também visando facilitar a marcação retórica. Com isso, o corpus contém, no total, 100 textos, sendo que cada texto tem, agora, tamanho máximo de 3 páginas, variando de 1 a 12Kb em tamanho de arquivo (totalizando 343Kb). A Tabela 1 mostra o número de textos e de palavras por área, indicando, também, a número médio de palavras por texto. O Gráfico 1 mostra a distribuição dos textos por área.

Tabela 1 – Número de textos por área

Área	**Número de textos**	**Número de palavras**	**Número médio de palavras por texto**
bd	7	2.859	408
es	33	16.084	487
h	29	16.980	585
ic	20	9.549	477
sd	11	7.172	652

Gráfico 1 – Distribuição dos textos por área



O padrão utilizado para nomeação dos 100 arquivos contendo os textos do corpus é o seguinte:

"Dissertacao_/Qualificacao_" + ÁREA + ID_TEXTO + "_parte" + ID_SEÇÃO

onde ÁREA é indicada pelas siglas das áreas da computação (bd, es, h, ic, sd ou sdpc), ID_TEXTO é um identificador único para o texto da área em questão e ID_SEÇÃO indica à qual seção da introdução o texto corresponde. Por exemplo, uma introdução de uma dissertação da área de banco de dados que contenha 2 seções

originaria dois arquivos chamados "Dissertacao_bd1_parte1.txt" e "Dissertacao_bd1_parte2.txt".

A seguir, apresenta-se uma breve introdução a *Rhetorical Structure Theory* – RST (Mann and Thompon, 1987), para, então, relatar a marcação retórica deste corpus.

3. *Rhetorical Structure Theory* – RST

A retórica é o meio pelo qual um texto é organizado para satisfazer o objetivo comunicativo que o escritor do texto pretendia atingir ao escrevê-lo. São as relações retóricas em um texto que delineiam como seu conteúdo é relacionado e como e em que medida cada unidade de conteúdo do texto, ou proposição, contribui para a satisfação das intenções do escritor. Diz-se, portanto, que as relações retóricas são responsáveis por atribuir coerência a um texto, relacionando suas partes e atribuindo-lhe sentido.

A RST estabelece relações retóricas que, de acordo com os autores, podem ser aplicadas a uma grande gama de textos sem necessitarem de modificações ou inclusões de outras relações. As relações são mostradas na primeira coluna da Tabela 2 (em inglês, como na obra de referência).

Tabela 2 – Relações retóricas da RST (Mann and Thompson, 1987)

Relações	**Multinuclear**	**Tipo das Relações**
ANTITHESIS	Não	*Presentational*
BACKGROUND	Não	*Presentational*
CIRCUMSTANCE	Não	*Subject Matter*
CONCESSION	Não	*Presentational*
CONDITION	Não	*Subject Matter*
ELABORATION	Não	*Subject Matter*
ENABLEMENT	Não	*Presentational*
EVALUATION	Não	*Subject Matter*
EVIDENCE	Não	*Presentational*
INTERPRETATION	Não	*Subject Matter*
JUSTIFY	Não	*Presentational*
MEANS	Não	*Subject Matter*
MOTIVATION	Não	*Presentational*
NON-VOLITIONAL CAUSE	Não	*Subject Matter*
NON-VOLITIONAL RESULT	Não	*Subject Matter*
OTHERWISE	Não	*Subject Matter*
PURPOSE	Não	*Subject Matter*
RESTATEMENT	Não	*Subject Matter*
SOLUTIONHOOD	Não	*Subject Matter*
SUMMARY	Não	*Subject Matter*
VOLITIONAL CAUSE	Não	*Subject Matter*
VOLITIONAL RESULT	Não	*Subject Matter*
CONTRAST	Sim	*Subject Matter*
JOINT	Sim	*Subject Matter*
LIST	Sim	*Subject Matter*
SEQUENCE	Sim	*Subject Matter*

Em casos padrões, essas relações se estabelecem entre duas proposições simples, geralmente expressas por segmentos adjacentes no texto, sendo uma nuclear (N) e outra complementar (S – "satélite"), indicando, respectivamente, a informação principal para a satisfação da intenção subjacente à relação e uma informação adicional, a qual influencia de alguma forma a interpretação que o leitor faz da informação nuclear. Quando ambas as informações relacionadas são igualmente importantes, diz-se que se tem uma relação multinuclear, isto é, com mais de um núcleo e nenhum satélite. Por exemplo, no trecho de texto "Embora você não goste, trabalhar é importante.", a proposição relativa à primeira oração é o satélite e a proposição relativa à segunda é o núcleo da relação retórica de oposição CONCESSION que existe entre as proposições. Por sua vez, no trecho de texto "O garoto chegou da escola e fez sua lição de casa. Depois, foi brincar com os amigos.", há uma relação SEQUENCE (indicando uma "seqüência" de eventos) entre as proposições relativas às orações "O garoto chegou da escola", "e fez sua lição de casa." e "Depois, foi brincar com os amigos.", sendo que todas são consideradas núcleos da relação, pois possuem a mesma importância. Na Tabela 2, somente as 4 últimas relações são multinucleares, como indicado na segunda coluna.

A definição de cada relação retórica pressupõe a existência de quatro tipos de informação necessários para determinar sua ocorrência entre duas proposições:

- Restrições sobre o núcleo (N);
- Restrições sobre o satélite (S);
- Restrições sobre a combinação do núcleo e do satélite (N+S);
- Efeito (ou intenção do escritor): especificação do efeito que a relação em questão causa no leitor, quando este interpreta um texto, ou do efeito pretendido pelo escritor, quando este seleciona tal relação para estruturar seu texto.

Como exemplo, as Figuras 1, 2, 3 e 4 mostram as definições das relações retóricas CONCESSION, BACKGROUND, EVIDENCE e ELABORATION.

Nome da relação: CONCESSION
Restrições sobre N: o escritor julga N válido. **Restrições sobre S**: o escritor não afirma que S pode não ser válido. **Restrições sobre N+S**: o escritor mostra uma incompatibilidade aparente ou em potencial entre N e S; o reconhecimento da compatibilidade entre N e S melhora a aceitação de N pelo leitor. **Efeito**: o leitor aceita melhor N

Figura 1 – Definição da relação retórica CONCESSION

Nome da relação: BACKGROUND
Restrições sobre N: o leitor não compreenderá suficientemente N antes de ler S. **Restrições sobre S**: sem restrições. **Restrições sobre N+S**: S aumenta a habilidade do leitor em compreender algum elemento em N. **Efeito**: a habilidade do leitor para compreender N aumenta.

Figura 2 – Definição da relação retórica BACKGROUND

Nome da relação: EVIDENCE
Restrições sobre N: o leitor poderia não acreditar em N de forma satisfatória para o escritor. **Restrições sobre S**: o leitor acredita em S ou o achará válido. **Restrições sobre N+S**: a compreensão de S pelo leitor aumenta sua convicção em N. **Efeito**: a convicção do leitor em N aumenta.

Figura 3 – Definição da relação retórica EVIDENCE

Nome da relação: ELABORATION
Restrições sobre N: sem restrições. **Restrições sobre S**: sem restrições. **Restrições sobre N+S**: S apresenta detalhes adicionais sobre a situação ou algum elemento de N. **Efeito**: o leitor reconhece S como apresentando detalhes adicionais sobre N.

Figura 4 – Definição da relação retórica ELABORATION

Como exemplo de análise, a Figura 6 mostra uma possível análise retórica para o texto da Figura 5 (cujos segmentos foram numerados para referência), fazendo uso de algumas das relações ilustradas anteriormente, onde os arcos horizontais indicam as proposições (expressas pelos segmentos textuais correspondentes) e os arcos verticais os núcleos das relações. O texto do exemplo faz parte do CorpusTCC.

Dependendo dos objetivos de pesquisa, as relações retóricas da RST podem ser organizadas de várias maneiras, observando-se, por exemplo, a participação do escritor na relação, a participação do leitor na relação, a ordem mais comum de realização do núcleo e do satélite no texto para cada relação (que pode ser núcleo antes do satélite ou satélite antes do núcleo), etc. A terceira coluna da Tabela 2 faz uma divisão em termos do efeito que as relações causam, que, neste trabalho, é a organização mais interessante de se analisar. Relações *presentational*, também chamadas de "intencionais" ou "pragmáticas", são aquelas que interferem em alguma inclinação do leitor, podendo aumentar seu desejo ou sua habilidade de realizar uma ação, crenças, aceitações e convicções; relações *subject matter*, também chamadas de "semânticas" ou "informativas", são aquelas cujo efeito pretendido é o de que o leitor apenas reconheça a existência da relação entre duas proposições, não interferindo em suas inclinações pessoais. A relação ELABORATION é um exemplo de relação *subject matter*, pois seu efeito pretendido é que o leitor reconheça que o satélite apresenta detalhes adicionais sobre o núcleo; a relação EVIDENCE, por sua vez, é uma relação *presentational* porque aumenta a convicção do leitor no núcleo da relação.

[1] A representação de grandes dicionários de língua natural, principalmente nos casos em que se trabalha com vários milhões (ou dezenas de milhões) de palavras, é um interessante problema computacional a ser tratado dentro da área de Processamento de Língua Natural. [2] Autômatos finitos, largamente usados na construção de compiladores, são excelentes estruturas para representação desses dicionários, [3] permitindo acesso direto às palavras e seus possíveis atributos.

[4] Um dicionário contendo mais de 430.000 palavras da língua portuguesa sem atributos, cuja representação em formato texto ocupa mais de 4.5Mb, pode ser convertido em um autômato compactado de apenas 218Kb.

Figura 5 – Texto 1

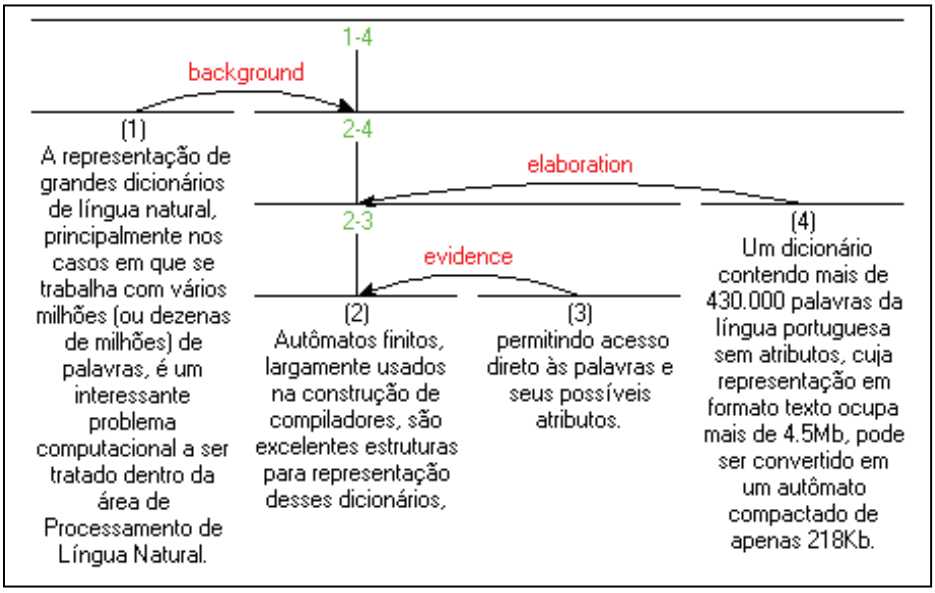


Figura 6 – Estrutura retórica do Texto 1

Vários autores modificaram/complementaram a RST para torná-la mais adequada para diversos fins, visto que a teoria apresenta alguns problemas de difícil solução. Por exemplo, como identificado pelos próprios autores da teoria, apesar da detalhada definição das relações, é difícil estabelecer uma única relação entre duas proposições, gerando, em muitos casos, ambigüidade. Outros autores (por exemplo, Moore e Pollack, 1992; Moser e Moore, 1996; Hovy, 1991, 1993) identificaram problemas relacionados à própria natureza das relações, afirmando que apenas as relações retóricas *presentational* são, de fato, retóricas; as relações *subject matter*, por sua vez, seriam de natureza semântica apenas. Marcu (1997, 2000), em especial, ao desenvolver o primeiro analisador retórico automático para o inglês, identificou e solucionou vários problemas (computacionais) que encontrou, os quais a RST não aborda, destacando-se:

- como determinar de forma consistente os segmentos textuais que correspondem a proposições simples, isto é, as unidades mínimas de significado;
- como montar a estrutura retórica completa de um texto a partir de suas estruturas retóricas parciais.

Para resolver a questão da segmentação, Marcu estabeleceu regras claras de segmentação (no nível da oração), fortemente ancoradas na sintaxe (vide Carlson e

Marcu, 2001). Para o segundo problema, Marcu sugeriu, inicialmente, que uma relação retórica qualquer R só poderia relacionar duas subestruturas retóricas T1 e T2, se R se estabelecesse entre as proposições mais nucleares dessas subestruturas. Com base nessa hipótese, Marcu desenvolveu algoritmos para, a partir das relações retóricas entre as proposições simples de um texto, derivar a estrutura retórica completa do texto.

Para a marcação retórica do CorpusTCC, as regras de segmentação de Marcu foram utilizadas. Apesar de terem sido desenvolvidas para o inglês, elas são genéricas o bastante para serem aplicadas ao português também. A sugestão de Marcu para montar estruturas retóricas completas a partir de estruturas parciais também foi observada durante a marcação do corpus, descrita a seguir[2].

4. Marcação Retórica do CorpusTCC

O CorpusTCC foi marcado retoricamente por somente uma pessoa[3] para manter a consistência da marcação e por questões de tempo. Pelos mesmos motivos, essa estratégia de marcação é recorrente na literatura (vide, por exemplo, Williams e Reiter, 2003).

Nas próximas subseções, são apresentados, respectivamente, a ferramenta utilizada para a marcação do corpus (subseção 4.1), a forma como a segmentação foi realizada (subseção 4.2), o elenco de relações utilizado para marcar o corpus (subseção 4.3), a estratégia de marcação (subseção 4.4) e alguns dados já extraídos do corpus marcado (subseção 4.5).

4.1. *RST Annotation Tool*

Para marcar os textos do CorpusTCC, foi utilizada a ferramenta *RST Annotation Tool* de Marcu[4], que é uma variação da ferramenta RSTTool de O'Donnel (1997). Por meio dessa ferramenta, é possível efetuar todos os passos da marcação retórica, ou seja, basicamente, segmentar o texto, escolher as relações retóricas e determinar o que é núcleo e satélite. A ferramenta também oferece possibilidade de escolher o elenco de relações retóricas a ser utilizado, desfazer operações e modificar estruturas retóricas já prontas, entre outras opções. Por fim, mas não menos importante, utilizando-se a ferramenta, é possível marcar um texto retoricamente por meio de diversas estratégias, conforme discutido por Carlson e Marcu (2001). A estratégia de marcação adotada para o CorpusTCC é discutida posteriormente. A Figura 7 mostra a interface da ferramenta. Na parte inferior da interface, o texto é apresentado ao usuário para segmentação; na parte superior, os segmentos identificados pelo usuário são exibidos, a partir dos quais o usuário constrói a estrutura retórica (na figura, o texto já está completamente segmentado e marcado retoricamente). O usuário pode alternar entre a segmentação e a marcação conforme queira.

Conforme o texto é segmentado e marcado retoricamente, a ferramenta armazena automaticamente os passos executados pelo usuário, permitindo que se

---

[2] Para mais detalhes sobre os trabalhos de Marcu e de outros autores que trabalharam com a RST, vide Pardo e Nunes (2003a). Para mais detalhes sobre segmentação textual, vide Pardo e Nunes (2003b).

[3] O primeiro autor deste relatório, que já tem experiência em trabalhar com a RST.

[4] http://www.isi.edu/~marcu/discourse/

analise o histórico de marcação de um texto, caso seja necessário. Além disso, diferentemente da ferramenta de O'Donnel, na ferramenta modificada de Marcu, ao final da marcação retórica, é oferecida ao usuário a possibilidade de 'salvar' a análise feita em arquivos de texto nos formatos SGML ou da linguagem de programação LISP.

A possibilidade de trabalhar com os textos visualmente, explorando toda a facilidade da manipulação gráfica, e salvar as análises feitas em arquivos de texto, o que facilita a manipulação computacional posterior, foi a grande motivação para a marcação do CorpusTCC pela *RST Annotation Tool*.

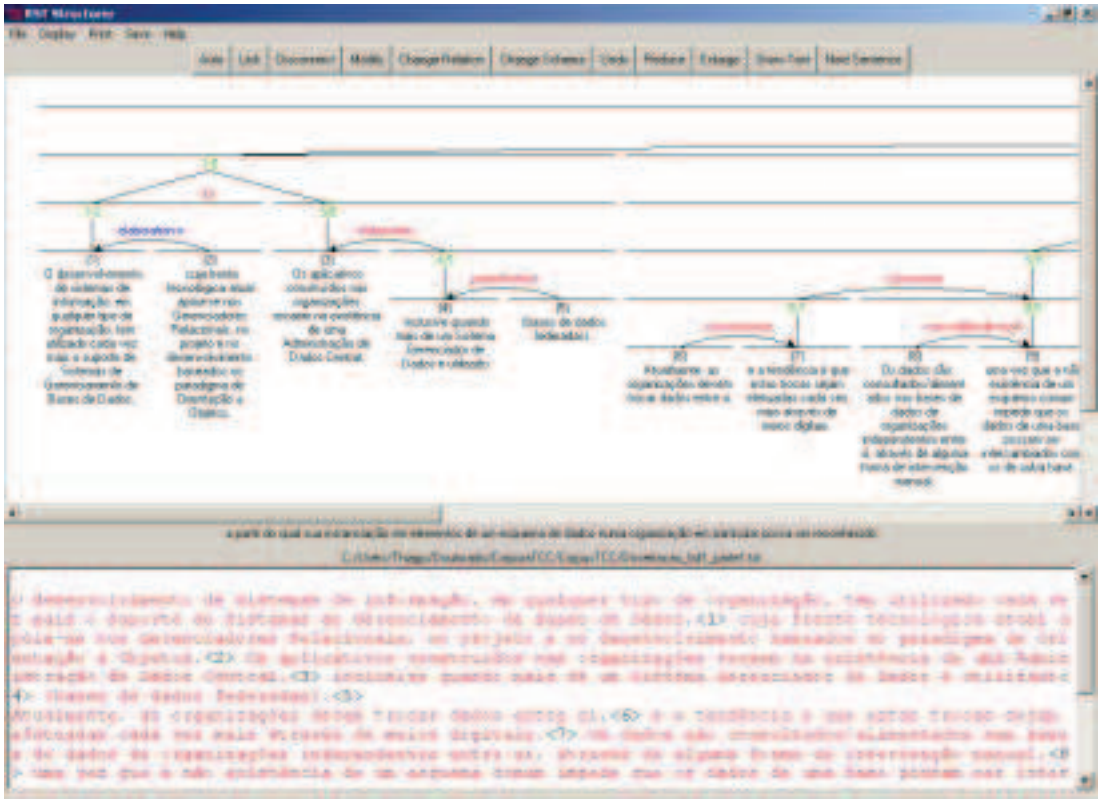


Figura 7 – Interface da *RST Annotation Tool*

4.2. Segmentação textual

De acordo com a RST, a segmentação de um texto pode se dar em qualquer nível, dependendo do tipo e da granularidade de análise que se quer realizar. Por exemplo, a segmentação pode se dar no nível da frase (de *phrase*, do inglês), da oração, da sentença, do parágrafo, etc. Para a marcação do CorpusTCC, seguindo o que se tem feito nos trabalhos recentes de análise retórica automática (Marcu, 1997, 2000; Corston-Oliver, 1998; Schilder, 2002), adotou-se, basicamente, a segmentação oracional. Isso se justifica pelo fato de que, normalmente, uma oração corresponde a uma proposição simples no texto.

Como comentado anteriormente, Carlson e Marcu (2001) desenvolveram regras consistentes para a segmentação textual oracional, regras estas fortemente ancoradas na sintaxe. Diz-se que essas regras são consistentes pelo fato de poderem ser aplicadas uniformemente a diferentes textos de maneira coerente e não ambígua. Para a segmentação do CorpusTCC, essas mesmas regras foram adotadas. Apesar de

terem sido desenvolvidas para o inglês, elas se mostraram genéricas o suficiente para serem aplicadas ao português também. Alguns exemplos de regras são:

1) orações principais são segmentos;
2) orações subordinadas com marcadores discursivos são segmentos;
3) orações coordenadas são segmentos distintos;
4) orações subordinadas substantivas subjetivas e objetivas não são segmentos;
5) orações relativas, apositivas e parênteses são considerados segmentos encaixados (*embedded*, no inglês).

Para o elenco completo das regras, vide a obra de referência. Vale comentar, entretanto, a decisão de se considerar segmentos encaixados na marcação retórica de um texto. Carlson e Marcu justificam essa escolha pelo fato dos segmentos encaixados serem abundantes nos textos. Não os considerar como segmentos implicaria perder muita granularidade na estruturação retórica de um texto. Para acomodar os segmentos encaixados na RST, algumas relações tiveram que ser incluídas no elenco de relações da RST, como discutido na subseção seguinte.

4.3. Elenco de relações retóricas

Inicialmente, para a marcação do CorpusTCC, utilizou-se o conjunto de relações proposto originalmente pela RST (Mann and Thompson, 1987). Conforme a marcação progrediu, percebeu-se a necessidade de mais algumas relações, que foram, então, extraídas do trabalho de Marcu (1997). A Tabela 3 mostra o elenco de relações utilizado. Como se pode notar, utilizaram-se tanto as relações *presentational* quanto as *subject matter*. Além disso, adotando-se a nomenclatura de Maier e Hovy (1991), algumas relações assumiram, aqui, um terceiro tipo, chamado *textual*. Essas relações são utilizadas para representar aspectos estruturais de um texto, possibilitando, também, acomodar os segmentos encaixados em uma estrutura retórica.

Tabela 3 – Elenco de relações retóricas para marcação do CorpusTCC

Relações	**Multinuclear**	**Tipo das relações**
ANTITHESIS	Não	*Presentational*
ATTRIBUTION	Não	*Textual*
BACKGROUND	Não	*Presentational*
CIRCUMSTANCE	Não	*Subject Matter*
COMPARISON	Não	*Subject Matter*
CONCESSION	Não	*Presentational*
CONCLUSION	Não	*Subject Matter*
CONDITION	Não	*Subject Matter*
ELABORATION	Não	*Subject Matter*
ENABLEMENT	Não	*Presentational*
EVALUATION	Não	*Subject Matter*
EVIDENCE	Não	*Presentational*
EXPLANATION	Não	*Subject Matter*
INTERPRETATION	Não	*Subject Matter*
JUSTIFY	Não	*Presentational*
MEANS	Não	*Subject Matter*
MOTIVATION	Não	*Presentational*

NON-VOLITIONAL CAUSE	Não	*Subject Matter*
NON-VOLITIONAL RESULT	Não	*Subject Matter*
OTHERWISE	Não	*Subject Matter*
PARENTHETICAL	Não	*Textual*
PURPOSE	Não	*Subject Matter*
RESTATEMENT	Não	*Subject Matter*
SOLUTIONHOOD	Não	*Subject Matter*
SUMMARY	Não	*Subject Matter*
VOLITIONAL CAUSE	Não	*Subject Matter*
VOLITIONAL RESULT	Não	*Subject Matter*
CONTRAST	Sim	*Subject Matter*
JOINT	Sim	*Subject Matter*
LIST	Sim	*Subject Matter*
SAME-UNIT	Sim	*Textual*
SEQUENCE	Sim	*Subject Matter*

Como exemplo de segmento encaixado e sua estruturação retórica, considere a sentença "A casa que a mulher comprou é bonita.", na qual a oração relativa "que a mulher comprou" é o segmento encaixado. Neste exemplo, a oração relativa especifica a casa a qual a sentença se refere, fragmentando a sentença "A casa é bonita". Para representar esses casos, em que há fragmentos que constituem uma única proposição, faz-se uso da relação multinuclear SAME-UNIT, como mostra a Figura 8.

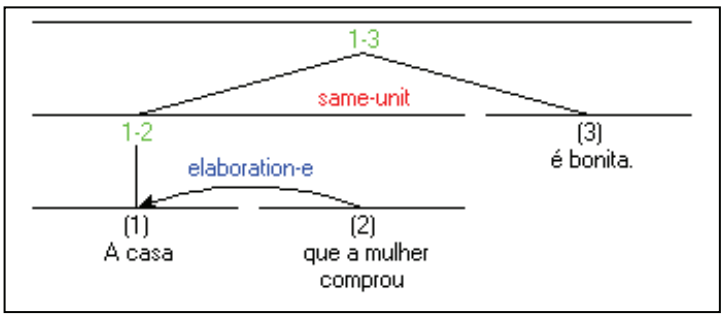


Figura 8 – Estrutura retórica com segmento encaixado

Na marcação do CorpusTCC, para diferenciar as relações retóricas que conectam proposições expressas pelos segmentos encaixados, adiciona-se "-e" ao final do nome dessas relações, como é possível notar na Figura 8. Essa é a mesma estratégia adotada por Marcu (1997, 2000). Neste relatório, para simplificação, essas relações serão referenciadas por relações encaixadas.

Como ilustração do uso das relações retóricas, é mostrado, nas Figuras 9-38, um exemplo de ocorrência de cada relação utilizada, exemplos estes retirados do próprio CorpusTCC, como recorte da interface da *RST Annotation Tool*. Quando necessário, algumas partes da estrutura retórica foram omitidas para que se pudesse focalizar a relação em questão. Para a definição das relações, sugere-se a leitura das obras de referência (Mann and Thompson, 1987; Marcu, 1997), as quais contêm as relações utilizadas neste trabalho.

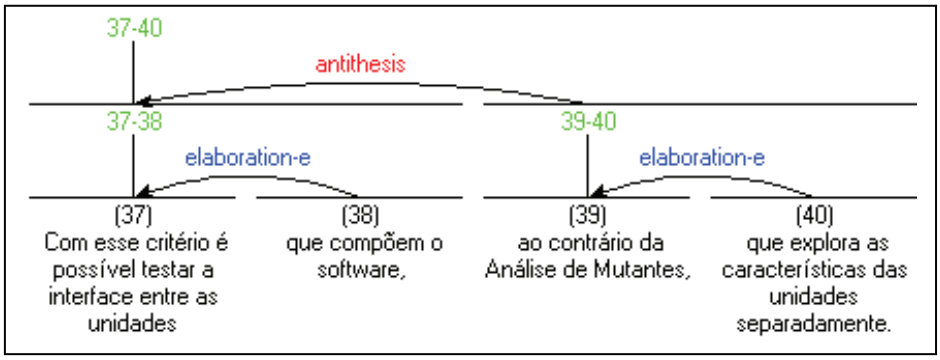


Figura 9 – Exemplo de ocorrência da relação ANTITHESIS

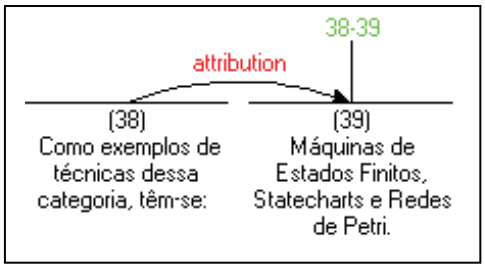


Figura 10 – Exemplo de ocorrência da relação ATTRIBUTION

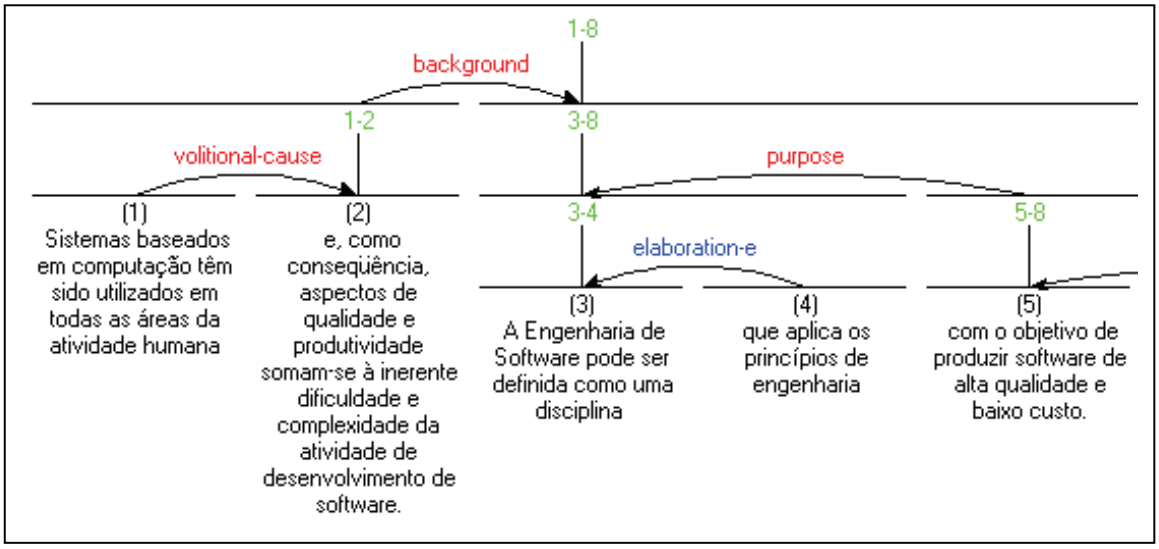


Figura 11 – Exemplo de ocorrência da relação BACKGROUND

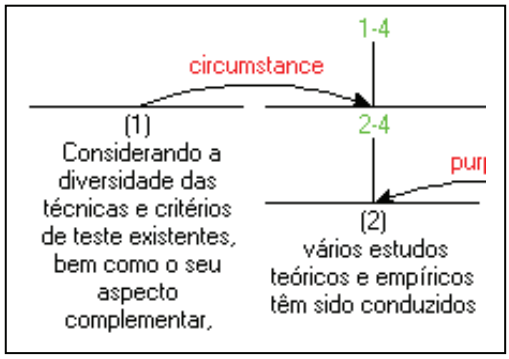


Figura 12 – Exemplo de ocorrência da relação CIRCUMSTANCE

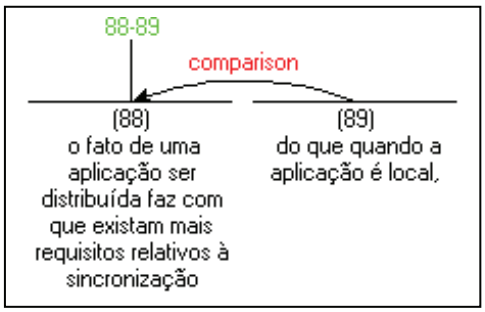


Figura 13 – Exemplo de ocorrência da relação COMPARISON

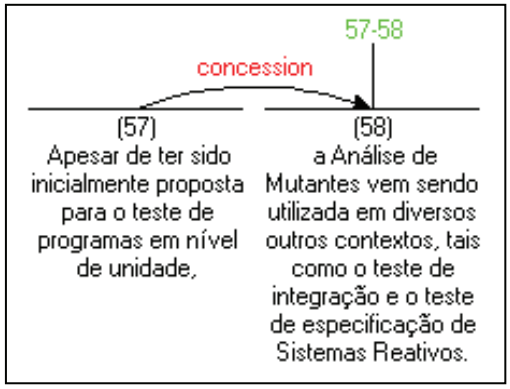


Figura 14 – Exemplo de ocorrência da relação CONCESSION

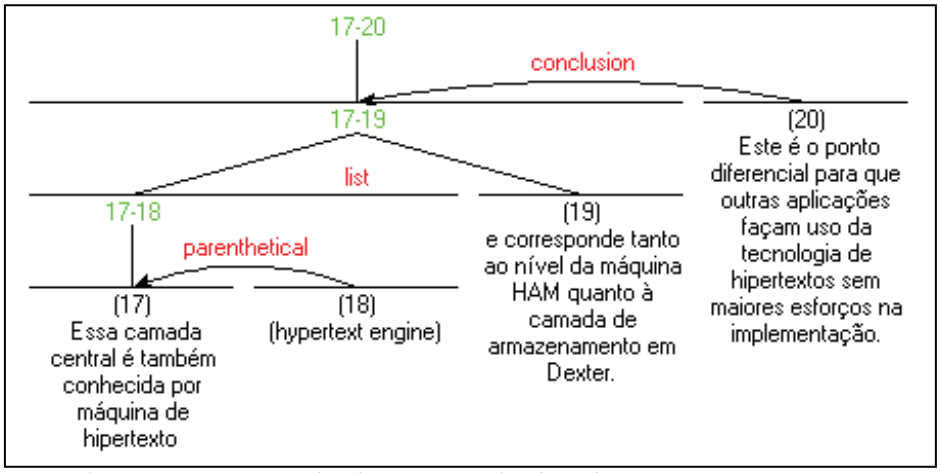


Figura 15 – Exemplo de ocorrência da relação CONCLUSION

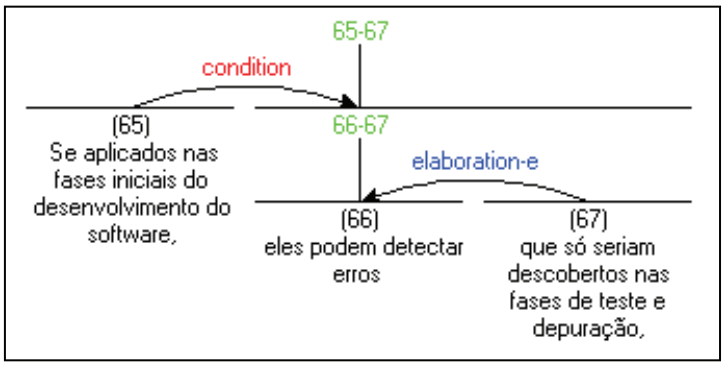


Figura 16 – Exemplo de ocorrência da relação CONDITION

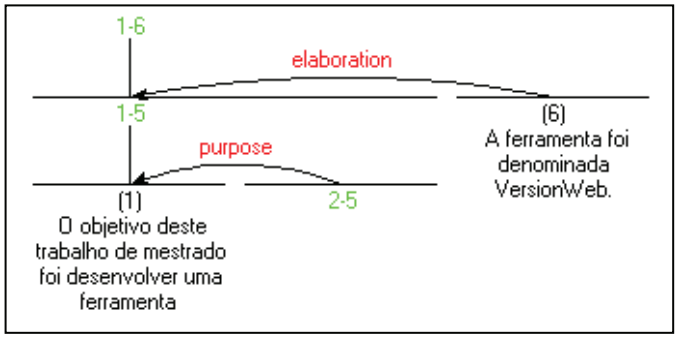


Figura 17 – Exemplo de ocorrência da relação ELABORATION

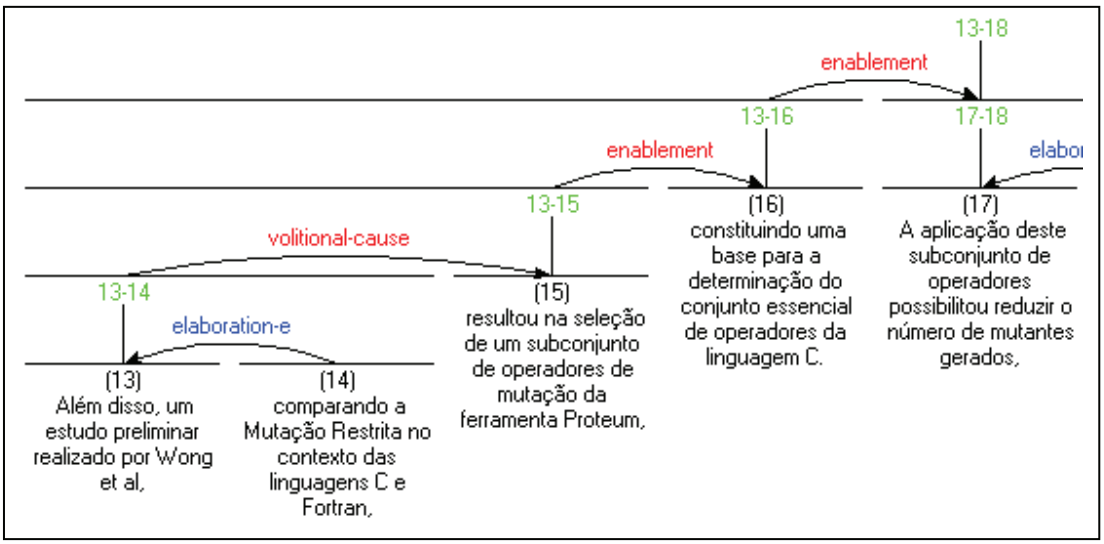


Figura 18 – Exemplo de ocorrência da relação ENABLEMENT

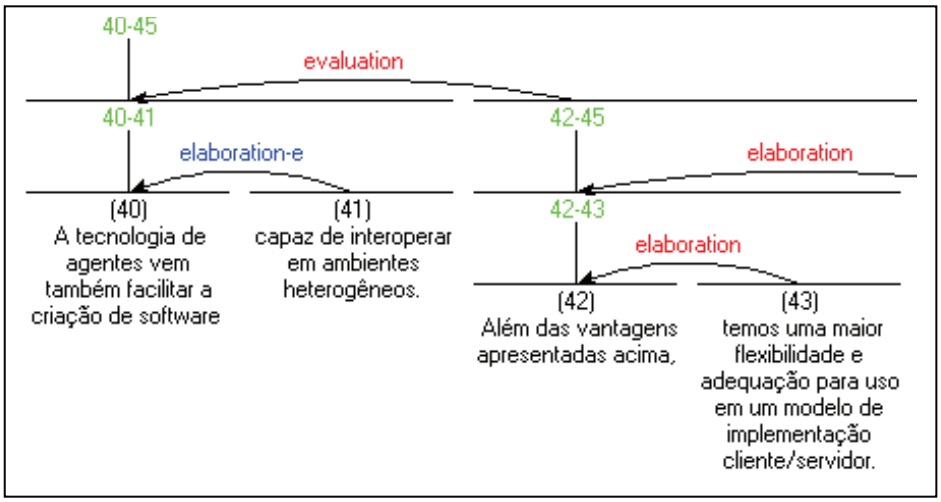


Figura 19 – Exemplo de ocorrência da relação EVALUATION

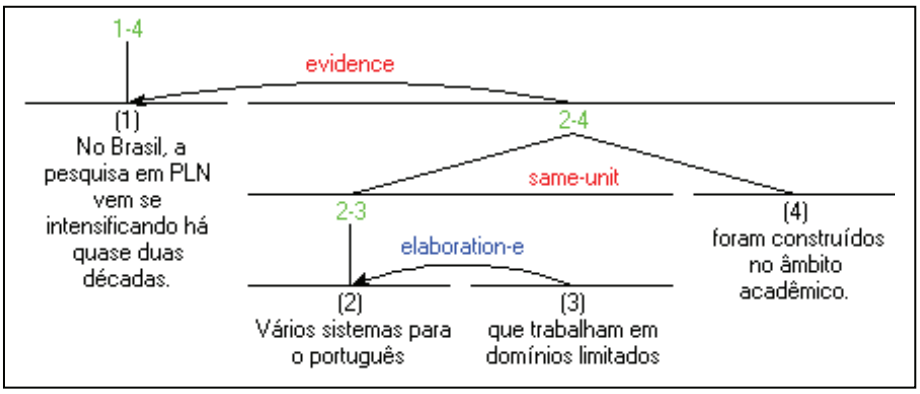


Figura 20 – Exemplo de ocorrência da relação EVIDENCE

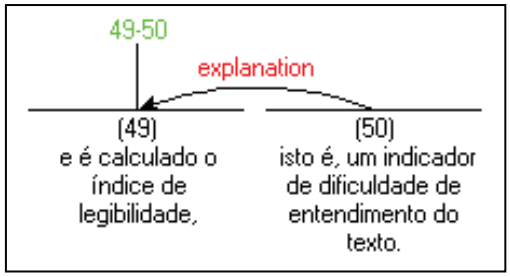


Figura 21 – Exemplo de ocorrência da relação EXPLANATION

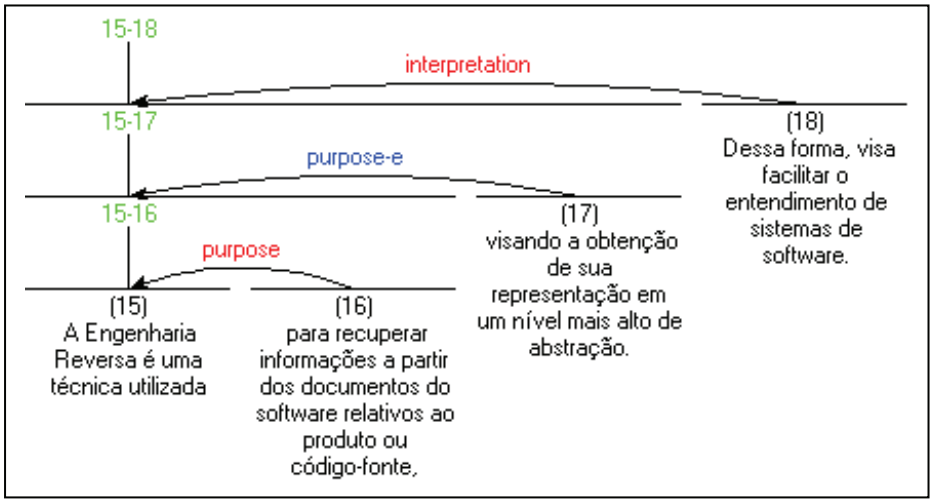


Figura 22 – Exemplo de ocorrência da relação INTERPRETATION

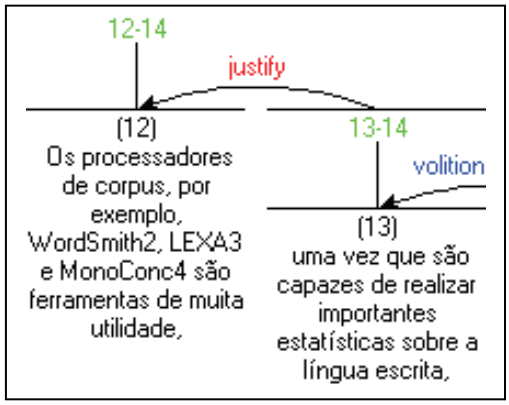


Figura 23 – Exemplo de ocorrência da relação JUSTIFY

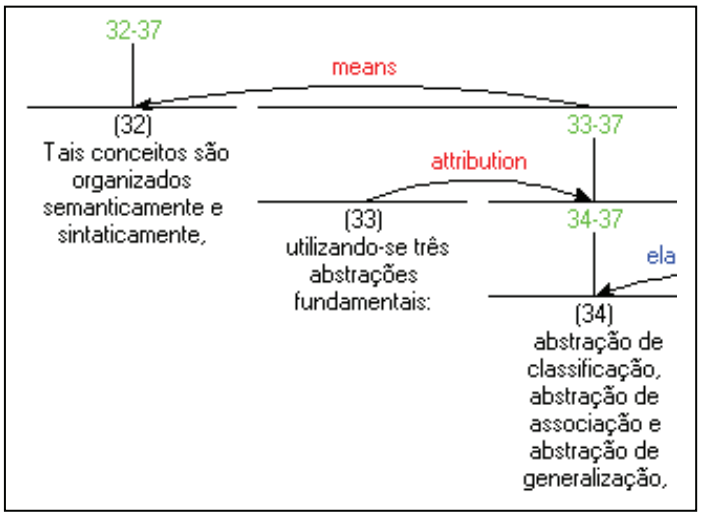


Figura 24 – Exemplo de ocorrência da relação MEANS

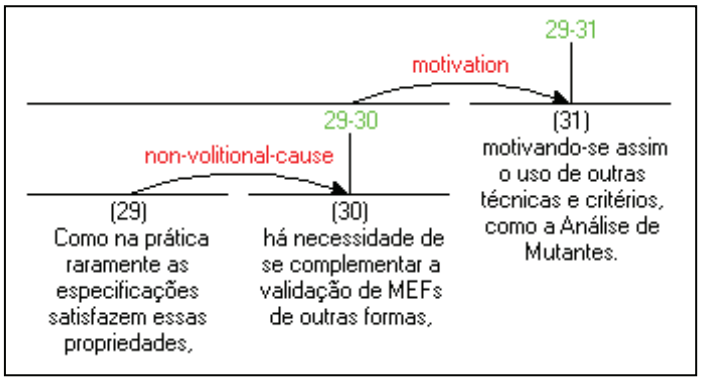


Figura 25 – Exemplo de ocorrência da relação MOTIVATION

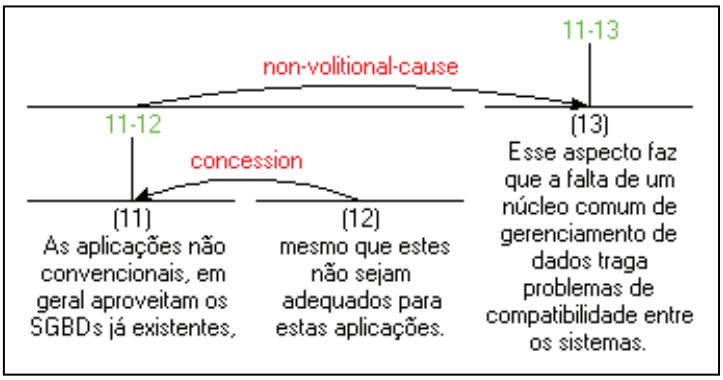


Figura 26 – Exemplo de ocorrência da relação NON-VOLITIONAL-CAUSE

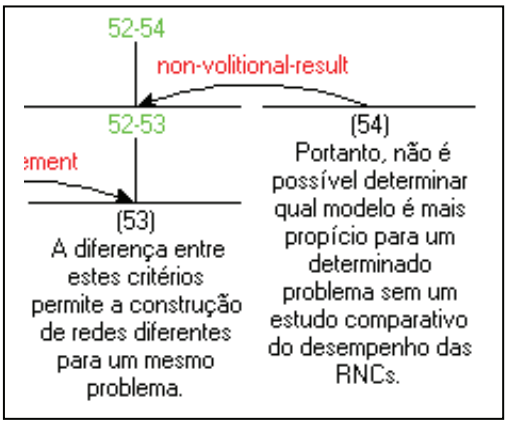


Figura 27 – Exemplo de ocorrência da relação NON-VOLITIONAL-RESULT

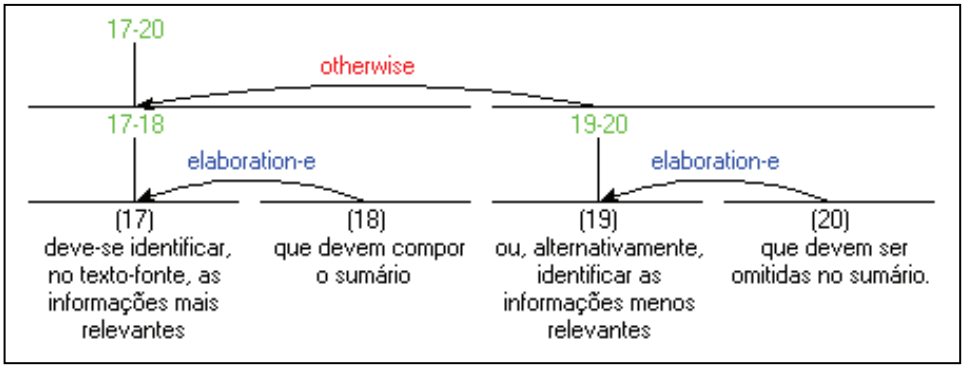


Figura 28 – Exemplo de ocorrência da relação OTHERWISE

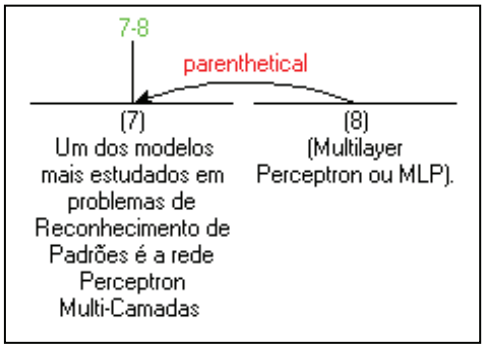


Figura 29 – Exemplo de ocorrência da relação PARENTHETICAL

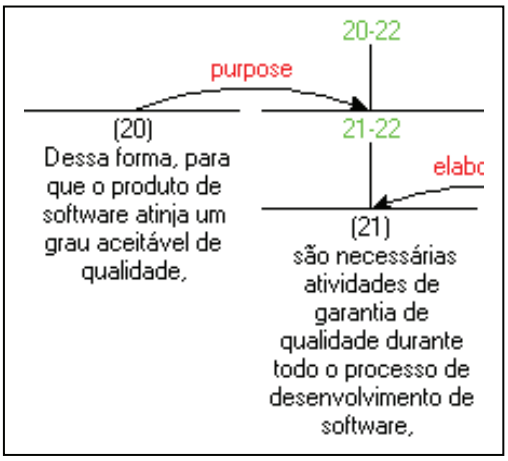


Figura 30 – Exemplo de ocorrência da relação PURPOSE

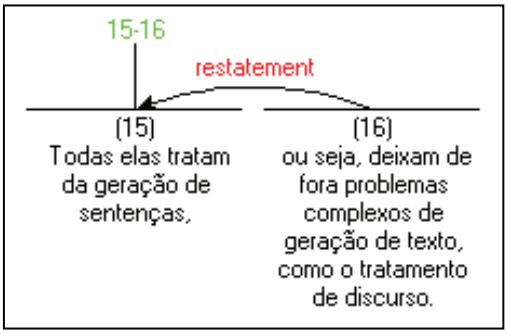


Figura 31 – Exemplo de ocorrência da relação RESTATEMENT

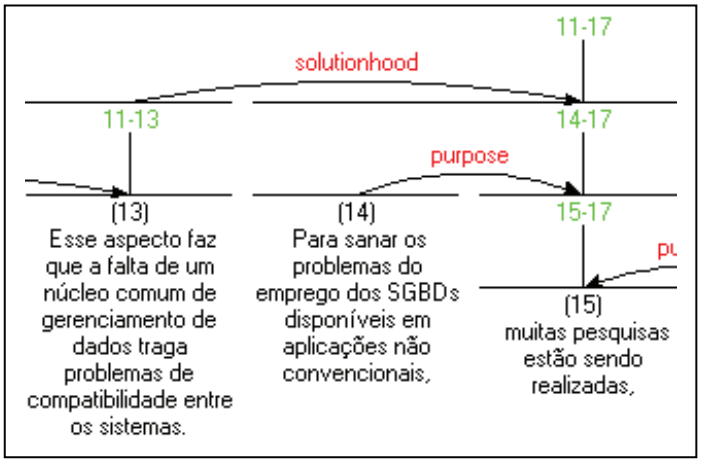


Figura 32 – Exemplo de ocorrência da relação SOLUTIONHOOD

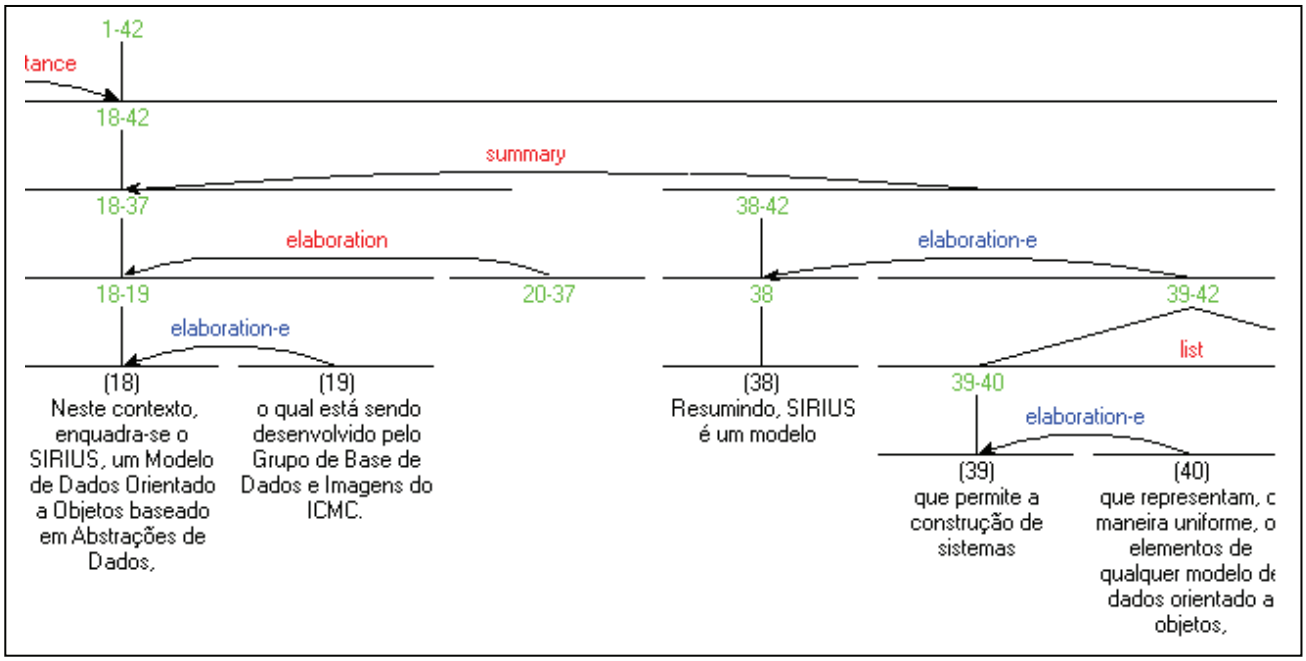


Figura 33 – Exemplo de ocorrência da relação SUMMARY

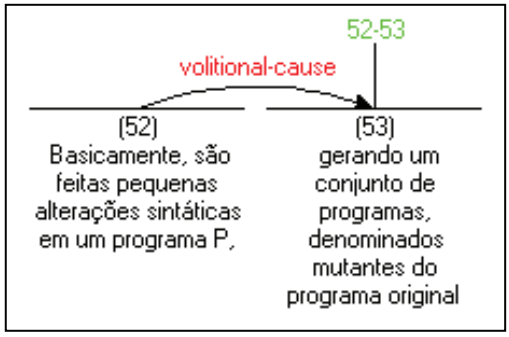


Figura 34 – Exemplo de ocorrência da relação VOLITIONAL-CAUSE

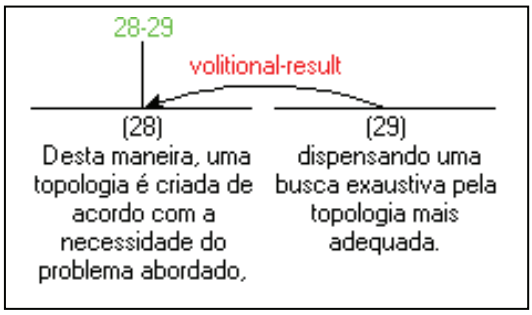


Figura 35 – Exemplo de ocorrência da relação VOLITIONAL-RESULT

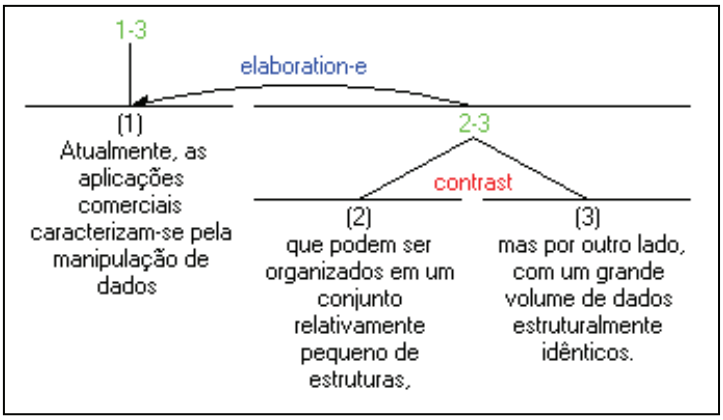


Figura 36 – Exemplo de ocorrência da relação CONTRAST

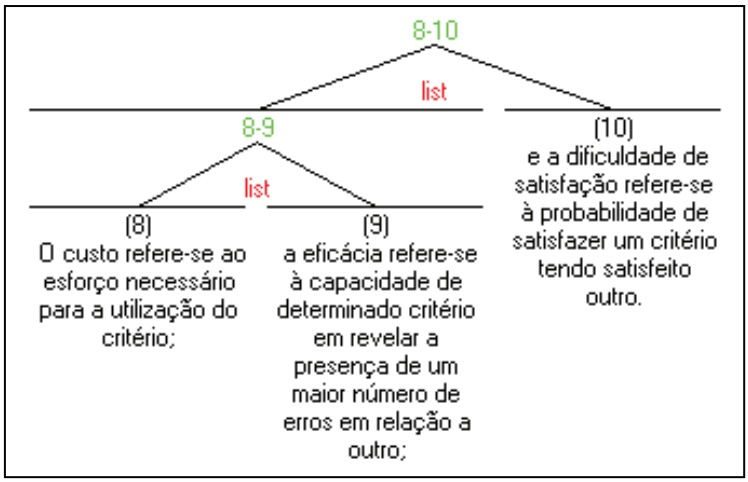


Figura 37 – Exemplo de ocorrência da relação LIST

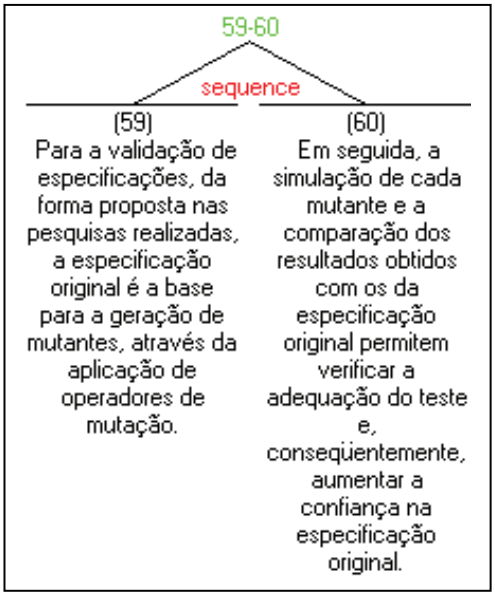


Figura 38 – Exemplo de ocorrência da relação SEQUENCE

4.4. Estratégia de marcação retórica

Carlson e Marcu (2001) discutem as estratégias possíveis para a marcação retórica de textos. Com base na observação dos anotadores, que eram pessoas treinadas para marcar textos retoricamente, perceberam diferentes estratégias de anotação: alguns liam o texto antes, outros não; alguns, conforme segmentavam o texto, já relacionavam a proposição correspondente à estrutura retórica parcial já construída, realizando o que foi chamado de marcação incremental; outros estruturavam os parágrafos do texto isoladamente, para então integrá-los em uma única estrutura.

Para a marcação do CorpusTCC, adotou-se a seguinte estratégia de anotação: primeiramente todas as proposições presentes em uma sentença eram relacionadas retoricamente; a seguir, todas as sentenças de um parágrafo eram relacionadas; por fim, os parágrafos eram relacionados. Essa estratégia se mostrou adequada e consistente para quase toda a marcação do corpus. Em algumas ocasiões, devido a estilos diferentes de escrita e, às vezes, problemas na estruturação textual, exceções ocorreram, de forma que, quando necessário, marcações incrementais eram realizadas.

Após a marcação do CorpusTCC estar completa, vários textos do corpus foram escolhidos aleatoriamente para terem sua marcação revisada. Em geral, não foram detectadas relações retóricas inadequadas. Em alguns casos, entretanto, notou-se que, em alguns contextos específicos, outras relações poderiam ser mais apropriadas. Nesses casos, a marcação não sofreu alteração, pois a ambigüidade retórica é natural e, para que o corpus seja realmente representativo para sua finalidade, o fato dessa ambigüidade estar representada no corpus é desejável.

No geral, pode-se dizer que os relatos de Carlson e Marcu (2001) sobre os problemas e questões da marcação retórica foram evidenciados durante a marcação do CorpusTCC. Nesses relatos, feitos a partir da observação de anotadores durante a marcação, os autores discutem questões como: a preferência dos anotadores por certas relações em casos de ambigüidade; a mudança no julgamento do que é adequado (em termos de escolha de relações retóricas) conforme os anotadores têm mais prática e experiência na marcação retórica; o tempo que os anotadores levam para marcar retoricamente textos conforme sua experiência no assunto aumenta. De fato, todos essas questões puderam ser observadas na marcação do CorpusTCC. Os pontos levantados por Carlson e Marcu podem explicar, também, porque algumas relações pareceram as mais apropriadas durante a marcação do CorpusTCC e, durante sua revisão, não pareceram.

4.5. Dados do corpus marcado

A Tabela 4 mostra o número de ocorrências e a freqüência das relações retóricas no CorpusTCC, não incluindo as relações encaixadas. Como se pode notar, algumas relações não ocorreram nenhuma vez (por exemplo, JOINT) ou muito poucas vezes (por exemplo, COMPARISON, OTHERWISE e SUMMARY). Por outro lado, como esperado, relações como ELABORATION e LIST foram as relações mais freqüentes, já que são as mais genéricas do elenco de relações. Marcu (1997) observou distribuições similares na marcação de seu corpus, apesar de ter utilizado um corpus de natureza diferente, composto por notícias jornalísticas.

A Tabela 5 mostra o número de ocorrências e a freqüência das relações retóricas encaixadas. Como se pode notar, muitas delas não ocorreram nenhuma vez.

Tabela 4 – Número de ocorrências e freqüências das relações retóricas

Relações	**Número de ocorrências**	**Freqüência (%)**
ANTITHESIS	21	0.53
ATTRIBUTION	185	4.69
BACKGROUND	112	2.84
CIRCUMSTANCE	121	3.07
COMPARISON	5	0.13
CONCESSION	62	1.57
CONCLUSION	14	0.35
CONDITION	19	0.48
ELABORATION	1030	26.10
ENABLEMENT	50	1.27
EXPLANATION	29	0.73
EVALUATION	10	0.25
EVIDENCE	15	0.38
INTERPRETATION	14	0.35
JUSTIFY	90	2.28
MEANS	48	1.22
MOTIVATION	18	0.46
NON-VOLITIONAL-CAUSE	64	1.62
NON-VOLITIONAL-RESULT	30	0.76
OTHERWISE	2	0.05
PARENTHETICAL	360	9.12
PURPOSE	318	8.06
RESTATEMENT	20	0.51
SOLUTIONHOOD	50	1.27
SUMMARY	4	0.10
VOLITIONAL-CAUSE	75	1.90
VOLITIONAL-RESULT	78	1.98
CONTRAST	89	2.26
JOINT	0	0,00
LIST	550	13.94
SAME-UNIT	393	9.96
SEQUENCE	70	1.77

Tabela 5 – Número de ocorrências e freqüências das relações encaixadas

Relações	**Número de ocorrências**	**Freqüência (%)**
ANTITHESIS-E	0	0,00
ATTRIBUTION-E	0	0,00
BACKGROUND-E	0	0,00
CIRCUMSTANCE-E	31	3,42
COMPARISON-E	6	0,66
CONCESSION-E	8	0,88

CONCLUSION-E	0	0,00
CONDITION-E	0	0,00
ELABORATION-E	651	71,85
ENABLEMENT-E	3	0,33
EXPLANATION-E	1	0,11
EVALUATION-E	5	0,55
EVIDENCE-E	0	0,00
INTERPRETATION-E	0	0,00
JUSTIFY-E	8	0,88
MEANS-E	18	1,99
MOTIVATION-E	1	0,11
NON-VOLITIONAL-CAUSE-E	2	0,22
NON-VOLITIONAL-RESULT-E	8	0,88
OTHERWISE-E	0	0,00
PURPOSE-E	139	15,34
RESTATEMENT-E	0	0,00
SOLUTIONHOOD-E	0	0,00
SUMMARY-E	0	0,00
VOLITIONAL-CAUSE-E	8	0,88
VOLITIONAL-RESULT-E	17	1,88

5. Considerações finais

Neste relatório, foram apresentadas a construção de um corpus de textos científicos da Computação, chamado CorpusTCC, e sua marcação retórica segundo a RST. Essa marcação consistiu na primeira das marcações necessárias para o desenvolvimento do DiZer, um analisador discursivo para o português do Brasil. Como próximos passos, o CorpusTCC deverá ser marcado semântica e intencionalmente.

Referências Bibliográficas

Carlson, L. and Marcu, D. (2001). *Discourse Tagging Reference Manual*. ISI Technical Report ISI-TR-545.

Corston-Oliver, S. (1998). *Computing Representations of the Structure of Written Discourse*. PhD Thesis, University of California, Santa Barbara, CA, USA.

Feltrim, V.D.; Nunes, M.G.V.; Aluísio, S.M. (2001). *Um corpus de textos científicos em Português para a análise da Estrutura Esquemática*. Série de Relatórios do NILC. NILC-TR-01-4.

Grosz, B. and Sidner, C. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, Vol. 12, N. 3.

Hovy, E. (1991). Approaches to the planning of coherent text. In C. Paris, W. Swartout and W. Mann (eds.), *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pp. 83-102. Kluwer Academic Publishers, Boston.

Hovy, E. (1993). In Defense of Syntax: Informational, Intentional, and Rhetorical Structures in Discourse. In the *Proceedings of the Intentionality and Structure in Discourse Relations Workshop*, pp. 35-39. Ohio, USA.

Jordan, M.P. (1992). An Integrated Three-Pronged Analysis of a Fund-Raising Letter. In W.C. Mann and S.A. Thompson (eds), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*, pp. 171-226.

Maier, E. and Hovy, E. H. (1991). A Metafunctionally Motivated Taxonomy for Discourse Structure Relations. In the *Proceedings of the 3rd European Workshop on Language Generation*. Innsbruck, Austria.

Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

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

Marcu, D. (2000). *The Theory e Practice of Discourse Parsing e Summarization*. The MIT Press. Cambridge, Massachusetts.

Moore, J. D. and Pollack, M. E. (1992). A problem for RST: the need for multi-level discourse analysis. *Computational Linguistics*, Vol. 18, N. 4, pp. 537-544.

Moser, M. and Moore, J. D. (1996). Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, Vol. 22, N. 3, pp. 409-419.

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

Pardo, T.A.S; Rino, L.H.M.; Nunes, M.G.V. (2003). DiZer: Uma Proposta de Análise Discursiva Automática para o Português do Brasil. In the *Anais do I Workshop em Tecnologia da Informação e da Linguagem Humana* - TIL. São Carlos-SP.

Pardo, T.A.S. e Nunes, M.G.V. (2003a). *Análise de Discurso: Teorias Discursivas e Aplicações em Processamento de Línguas Naturais*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação - ICMC, Universidade de São Paulo, no. 196.

Pardo, T.A.S. e Nunes, M.G.V. (2003b). *Segmentação Textual Automática: Uma Revisão Bibliográfica*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação - ICMC, Universidade de São Paulo, no. 185.

Pinheiro, G.M. e Aluísio, S.M. (2003). *Corpus NILC: Descrição e Análise Crítica com Vistas ao Projeto Lacio-Web*. Série de Relatórios Técnicos do Instituto de Ciencias Matemáticas e de Computação – ICMC, Universidade de São Paulo, N. 190.

Schilder, F. (2002). Robust discourse parsing via discourse markers, topicality and position. In J. Tait, B.K. Boguraev and C. Jacquemin (eds.), *Natural Language Engineering*, Vol. 8. Cambridge University Press.

Williams, S. and Reiter, E. (2003). A corpus analysis of discourse relations for Natural Language Generation. In the *Proceedings of Corpus Linguistics*, pp. 899-908. Lancaster University.