Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática

Sandra Collovini[1], Thiago I. Carbonel[2], Juliana Thiesen Fuchs[1], Jorge César Coelho[1], Lúcia Rino[2], Renata Vieira[1]

[1]Universidade do Vale do Rio dos Sinos (UNISINOS)
Av. Unisinos 950 – 93022-000 – São Leopoldo – RS – Brasil

[2]Universidade Federal de São Carlos (UFSCar)
Rodovia Washington Luís (SP-310) – 13565-905 – São Carlos - SP – Brasil

scollovini@turing.unisinos.br, {julianatf,ccoelho}@icaro.unisinos.br

renatav@unisinos.br, {thiagocarbonel,lucia}@dc.ufscar.br

Abstract. *This paper presents the Summ-it corpus, built to provide a basis for the study of discourse summarization along with the phenomena of anaphoric and rethorical relations.*

Resumo. *Este artigo apresenta o corpus Summ-it, elaborado com o objetivo de embasar pesquisas de discurso envolvendo relações anafóricas e retóricas e a sumarização automática.*

1. Introdução

Este artigo apresenta o desenvolvimento do corpus Summ-it e a sua preparação para pesquisas sobre o discurso e a Sumarização Automática (SA) de textos em português no âmbito dos projetos ProCaCoSA (Processamento de Cadeias de Co-referência para a Sumarização Automática de textos em Português) e PLN-BR (Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil)[1]. O problema em foco é a *quebra de cadeias de correferência* (CCRs), ou seja, pelo processo de SA, um segmento textual contendo uma anáfora é incluído no sumário, mas o segmento que contém seu antecedente não é inserido. A construção do corpus Summ-it anotado com informações discursivas visa fornecer subsídios para enriquecer os modelos de SA a fim de melhorar a coerência e o grau de informatividade dos sumários automáticos. Consideramos, neste artigo, o uso de informação de correferência e o emprego de relações retóricas descritas na Teoria RST - *Rhetorical Structure Theory* [Mann and Thompson 1988]. Descrevemos aqui os processos de anotação de correferência e de relações retóricas do corpus Summ-it. O primeiro processo busca identificar as entidades do discurso (por exemplo, sintagmas nominais) mencionadas e/ou retomadas em um texto. O segundo permite a estruturação de um texto relacionando suas unidades discursivas por meio das relações RST, resultando em uma estrutura arbórea, chamada árvore RST. O corpus Summ-it constitui-se de 50 textos jornalísticos do caderno de Ciências da Folha de São Paulo retirados do corpus PLN-BR. Cada documento corresponde a um arquivo texto com tamanho entre 1 Kbytes e 4 Kbytes (de 127 a 654 palavras).

[1]www.inf.unisinos.br/˜ renata/laboratorio/projetos.htm

No que segue, introduzimos a metodologia da anotação de correferência do corpus Summ-it na Seção 2. Na Seção 3, apresentamos sua anotação retórica por meio de relações RST. Na seção 4, apresentamos resultados preliminares das anotações. Por fim, na Seção 5, apresentamos as considerações finais.

2. Anotação de correferência

Tradicionalmente, o processo de resolução de correferência busca mapear relações de identidade (fazem referência a uma mesma entidade discursiva) ou de dependência (tais como, relações de posse) entre expressões de um texto. Particularmente, nosso trabalho, centraliza-se nas expressões nominais (núcleo nome ou pronome). Com efeito, a organização dessas expressões é importante tanto para interpretação quanto para produção textual - aspectos fundamentais num contexto de sumarização. Basicamente, nesse processo de organização, utilizamos termos/expressões que retomam outros termos do próprio texto, constituindo, assim, cadeias correferenciais. Para exemplificar, segue o trecho de um texto com suas expressões correferentes. No exemplo, podemos observar que a expressão *os pesquisadores* retoma *Pesquisadores do Museu Nacional do Rio de Janeiro*, ou seja, *Pesquisadores do Museu Nacional do Rio de Janeiro* é antecedente de *os pesquisadores*. Assim, as duas expressões formam uma cadeia de correferência (uma cadeia devendo possuir pelo menos duas expressões referenciais).

- *Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram ontem a descoberta de uma nova espécie de dinossauro ...* ***os pesquisadores*** *conseguiram estimar que o bicho fosse um filhote de 1,5 metro de altura.*

Na literatura, encontramos vários trabalhos que apresentam sistemas de resolução de correferência automáticos [Müller et al. 2002, Ng and Cardie 2002, Poesio et al. 2005]. Para a avaliação desses sistemas é necessário um corpus padrão, geralmente, anotado de modo manual. Para o inglês são utilizados os corpora MUC-6 e ACE, disponibilizados pelo LDC (*Linguistic Data Consortium*)[2]. Para o português, não existe ainda um corpus anotado com esse tipo de informação. Para preencher essa lacuna propomos a anotação do corpus Summ-it com informações discursivas, cuja metodologia é descrita a seguir.

2.1. Metodologia de anotação

A anotação de correferência manual de nosso corpus seguiu instruções para a anotação de informações de correferência e de referências dêiticas, designadamente, elaboradas para o discurso escrito do português [Coelho et al. 2006]. A metodologia de anotação é baseada em estudos realizados pelos projetos ANACORT[3], TeXto[4] e VENEX[5] e conta com o uso do analisador sintático do Português PALAVRAS [Bick 2000] e da ferramenta de anotação MMAX (*Multi-Modal Annotation in XML*) [Müller and Strube 2001].

A anotação seguiu várias etapas: seleção das unidades de interesse, denominadas *markables*, identificação de suas configurações morfossintáticas, indicação das relações entre os diversos *markables*, classificação dos mesmos e classificação dos relacionamentos anafóricos correferenciais e associativos.

[2]http://www.ldc.upenn.edu/
[3]www.inf.unisinos.br/˜ renata/laboratorio/anacort_descricao.htm
[4]www.inf.unisinos.br/˜ renata/laboratorio/texto_index.htm
[5]cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf

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

Tabela 1. Atributos dos markables

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

Seleção das unidades de interesse - *markables*: São os sintagmas nominais (SNs) que têm como núcleo um nome comum (*os pesquisadores*), um nome próprio (*o Museu Nacional*) ou um pronome (*Eles*). Esta etapa foi realizada de forma semi-automática. Primeiramente, os SNs foram extraídos automaticamente, com base nas informações do PALAVRAS. Após, os *markables* foram revisados manualmente utilizando a MMAX, seguindo as instruções detalhadas em [Coelho et al. 2006].

Identificação das configurações morfossintáticas dos *markables*: As configurações morfossintáticas são descritas pelos atributos *np_form* e *pro_form* (veja Tabela 1), para distinguir os SNs com núcleo nome dos pronomes respectivamente. As possíveis configurações para os sintagmas nominais com núcleo nome são:

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

Para os pronomes temos como configurações:

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

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

pelo atributo *pointer* da MMAX.

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

- ***new***: novo referente no discurso que não apresenta parte de seu sentido ancorado em uma expressão anterior (*o nordeste brasileiro*).
- ***old***: a expressão retoma um referente já introduzido por uma expressão anterior (*o gambá - o animal*).
- ***associative***: introduz um novo referente no discurso, mas cujo significado está ancorado em uma expressão anterior (*macieiras - a maça*).
- ***deictic***: a informação requerida para interpretação da expressão não é encontrada no texto, mas na situação comunicativa (*a semana passada*).

Classificação dos relacionamentos anafóricos correferenciais: Neste caso, temos uma subclassificação de *markables* com *status=old* e atributo *is_anaphoric* (ver Tabela 1) em:

- ***direct***: a expressão tem um antecedente que apresenta nome núcleo idêntico (*a carta - uma carta*).
- ***indirect***: a expressão tem um antecedente que apresenta núcleo diferente (*a carta - o documento*).
- ***encapsulation***: a expressão retoma um trecho de texto maior que um sintagma, por exemplo (*a operação* retoma a sentença *O Banco Central interveio ontem para segurar a cotação do dólar*). Aqui utilizamos o atributo *comment* (veja Tabela 1).

Classificação dos relacionamentos anafóricos associativos: De forma análoga à anterior, os *markables* em foco aqui são aqueles com atributo *is_bridging*, que permitem a subclassificação dos *markables* do tipo *associative* nas relações seguintes (segundo as diretrizes adotadas no projeto VENEX[6]:

- ***element-of***: a expressão anafórica é um elemento de um grupo previamente introduzido (*algumas áreas - a área*). Quando o elemento ocorre antes do conjunto, deve-se usar a relação inversa: element-of-inv (*o único dente, um molar inferior - os molares*).
- ***subset-of***: a expressão anafórica refere-se a um subconjunto de uma entidade introduzida anteriormente no texto (*os bichos - os machos*).
- ***part-of***: a expressão invoca parte de uma entidade já mencionado (*macieiras - a maçã*). Quando a parte ocorre antes do todo, deve-se usar a relação inversa: part-of-inv (*São Paulo - o país*).
- ***entity-attribute***: a expressão refere-se a um atributo de uma entidade previamente mencionada (*uma pesquisa com 240 casais - os resultados*).
- ***possessor-thing***: o antecedente possui a entidade evocada pela expressão associativa (*a superativação do gene - os seus efeitos colaterais*).
- ***other-brigding***: outros tipos de relação não definidos pelos anteriores (*o rio - a correnteza*).

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

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

2.1.1. Exemplo de anotação de correferência

O exemplo apresentado aqui corresponde a alguns segmentos de um texto do corpus Summ-it. As descrições definidas são apresentadas em negrito no exemplo (correspondendo ao atributo *text* nas figuras abaixo) e seus antecedentes são sublinhados.

- *O estudo, realizado por pesquisadores do Imperial College, em Londres, analisa células-tronco da medula óssea (...)* ***A pesquisa*** *possibilitará que pessoas com dano no fígado usem as suas próprias células-tronco d****a medula óssea*** *para produzir células hepáticas.*

No exemplo, observamos pela progressão do texto que a expressão *a pesquisa* retoma a expressão *o estudo*, ambas referem-se à mesma entidade com nome núcleos diferentes (*pesquisa - estudo*), portanto, *status=old* e *is_anaphoric=indirect* - Figura 1(a). Já a expressão *a medula óssea* retoma a expressão anterior de mesmo nome núcleo *a medula óssea*. Neste caso, indicamos o *status=old* e *is_anaphoric=direct* - Figura 1(b).

Figura 1. *Markables* referentes ao exemplo de anotação do corpus Summ-it

```
(a)<markable id="markables_31"
text="a pesquisa"
span="word_97..word_98"
is_anaphoric="indirect"
np_n="yes" np_form="def-np"
status="old"/>
member="set_23"/>
```

```
(b)<markable id="markables_56"
text="a medula óssea"
span="word_173..word_174"
is_anaphoric="direct"
np_n="yes"  np_form="def-np"
status="old"
member="set_12"/>
```

3. Anotação de relações RST

A Teoria RST procura descrever a estrutura textual relacionando segmentos de discurso, na prática representados por segmentos textuais, por meio das chamadas relações RST. Para a estruturação textual, é fundamental a noção de coerência textual, sendo que cada relação RST atribui um papel específico a cada segmento de texto envolvido. As relações RST podem ser mononucleares - quando se estabelecem entre segmentos mais centrais (os núcleos) e segmentos ditos complementares, ou periféricos à informação contida no núcleo (os satélites) - ou multinucleares - quando se estabelecem entre segmentos com teor de informatividade similar. Cada relação é definida segundo dois critérios: o de condições (ou restrições) de aplicabilidade da relação a um dado contexto textual e o de efeito que a relação indicada pode causar sobre o leitor, em função dos objetivos de produção textual do escritor, isto é, de suas opções de organização e apresentação. É nesse sentido que a RST é ”retórica” [Mann et al. 1992].

A relação BACKGROUND, por exemplo, é descrita da seguinte forma [Mann et al. 1992][7]:

- **Nome da relação**: BACKGROUND
- **Condições no núcleo (N)**: o leitor não compreenderá o núcleo suficientemente antes de ler o texto do satélite.
- **Condições na combinação núcleo-satélite (N+S)**: o satélite aumenta a capacidade do leitor para compreender um elemento no núcleo.
- **Efeito**: a capacidade do leitor para entender o núcleo é aumentada.
- **Locus do efeito**: N.

[7]Toda relação RST possui *template* similar a este, isto é, mesmos campos e, portanto, descrição funcional com estrutura similar. A tradução das definições originais das relações para o português é nossa.

Descrições funcionais como essa servem para especificar julgamentos particulares de plausibilidade, quer para a produção, quer para a interpretação textual. No nosso caso de análise RST visando à anotação dos textos do corpus Summ-it, a determinação de uma certa relação RST entre dois segmentos textuais pelo analista deve se basear nessas descrições e, portanto, deve considerar as condições e o efeito de cada relação. Na Figura 2, é possível visualizar a relação BACKGROUND aplicada a um trecho de texto do corpus Summ-it.

3.1. Metodologia de anotação

Dois analistas, especialistas em RST, estruturaram cada texto de nosso corpus com o apoio da ferramenta de suporte à anotação RSTTool [O'Donnell 2000]. Essa ferramenta fornece recursos gráficos úteis para a visualização da estrutura arbórea prevista na Teoria RST e não automatiza a análise em nenhum aspecto. A Figura 2 é construída com apoio dessa ferramenta. Como se pode notar, cada segmento de uma relação é marcado por uma linha horizontal. O segmento nuclear é marcado por uma linha vertical. A relação é marcada por uma seta que aponta sempre do satélite para o núcleo, exceto no caso de relações multinucleares.

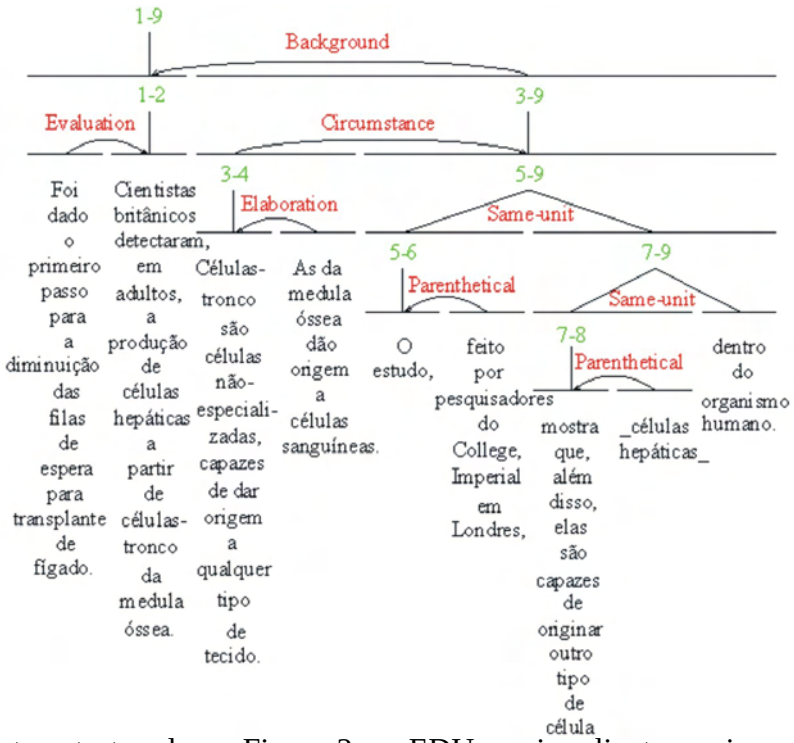
A anotação seguiu os critérios de [Mann and Thompson 1988] e [Carlson and Marcu 2001], além de algumas diretrizes adicionais elaboradas pelos próprios anotadores para dirimir dúvidas. Para identificar as relações em um texto, Mann e Thompson (2001) sugerem que o analista deve primeiramente segmentá-lo em unidades textuais ou, segundo [Carlson and Marcu 2001], unidades discursivas elementares, aqui representadas pela sua sigla em inglês, EDUs (*Elementary Discourse Units*). Uma EDU é considerada um bloco mínimo de construção de uma árvore discursiva (sempre ocorrendo como uma folha da árvore), correspondendo a uma proposição elementar, no nível discursivo. O tamanho da unidade de discurso é arbitrário para a RST, podendo abranger desde itens lexicais típicos até parágrafos inteiros ou unidades ainda maiores. Porém, as unidades devem ter integridade funcional independente. Carlson e Marcu (2001) sugerem que se considere a oração como a unidade elementar do discurso, usando indícios lexicais e sintáticos para ajudar na determinação de fronteiras. Após delimitar EDUs no texto, o passo seguinte consiste em estabelecer as relações entre elas. Conforme [Carlson and Marcu 2001], as EDUs são ligadas por meio das relações RST, o que cria uma estrutura hierárquica.

A RSTTool permite a seleção de um conjunto pré-definido de relações RST para análises. Vários conjuntos já existem no pacote da ferramenta, definidos por diversos pesquisadores. Novos conjuntos também podem ser adicionados, envolvendo a redefinição de relações já existentes ou a conversão de relações em conjuntos de relações mais precisamente definidos. Para realizar a anotação do corpus Summ-it, os anotadores usaram o conjunto de 32 relações apresentado em [Pardo 2005], que reúne 26 relações do conjunto de [Mann and Thompson 1988] e seis relações do conjunto de [Carlson and Marcu 2001]. O conjunto de [Pardo 2005] foi adotado por ser o utilizado no analisador discursivo automático DiZer [Pardo 2005], 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 seus respectivos indicadores, ou seja, marcadores discursivos e expressões indicativas.

3.1.1. Exemplo de anotação de relações RST

Como indica a Figura 2, a estrutura RST de um texto assinala EDUs de modo a identificar as unidades nucleares e as não-nucleares (satélites). Essa característica permite, portanto, identificar a relevância das unidades do discurso de acordo com a sua nuclearidade e, assim, selecionar as unidades mais importantes para a composição de uma versão condensada do texto, ou seja, o resumo. Esta proposta já foi explorada de diferentes formas, em diversos trabalhos para diferentes línguas, por exemplo, [Ono et al. 1994, Marcu 1997, Marcu 1999, Marcu 2000, O'Donnell 2000, Rino 1996]. Marcu, particularmente, usa a idéia de saliência decorrente da nuclearidade para determinar, na árvore RST, o grau de saliência de seus segmentos. A partir da ordenação dos graus de cada segmento do discurso (ou componente da árvore RST), seu sumarizador organiza o sumário de acordo com a taxa de compressão, selecionando os segmentos mais salientes.

Figura 2. Análise de parte do texto CIENCIA_2000_17109, diagramada na RSTTool



No texto estruturado na Figura 2, as EDUs mais salientes seriam a EDU 2 [*Cientistas britânicos detectaram, em adultos, a produção de células hepáticas a partir de células-tronco da medula óssea*] e a EDU 3 [*Células-tronco são células não-especializadas, capazes de dar origem a qualquer tipo de tecido*]. Sua justaposição permitirá, assim, construir uma versão condensada, informativa e, nesse caso, textual, do conteúdo mais relevante explicitado no texto-fonte[8].

4. Resultados preliminares da anotação do corpus Summ-it

Os resultados da anotação de correferência apresentados aqui estão baseados no cálculo da média entre a anotação de dois anotadores para cada texto. Para facilitar a anotação dos

[8]Neste caso, a SA seria extrativa, pois o resultado seria exatamente um *extrato* desse texto.

50 textos que constituem o corpus Summ-it, o mesmo foi dividido em 4 partes. A anotação de RST e de correferência foi concluída para as 4 partes, sendo que o consenso final da anotação de correferência está em fase de conclusão. Cabe salientar que, na anotação de correferência, os anotadores indicaram as relações entre os *markables* (atributos *pointer* e *member*) para todas as configurações dos SNs. E, na anotação das CCRs (atributo *member*), os anotadores identificaram um total de 560 CCRs no corpus Summ-it, tendo cada CCR uma média de 3 membros e contendo a, mais extensa, 16 membros.

Apresentamos aqui, primeiramente, a distribuição das configurações morfossintáticas dos SNs do corpus Summ-it (Tabela 2). Podemos observar que de 5047 *markables*, a maior parte corresponde aos SNs com nome núcleo (95,18%), pronomes sendo somente 4,82%. Devido a isso, concentramos nossa atenção somente nos SNs, mais particularmente nas descrições definidas (*np_form*=def-np e *np_form*=def-pn). A etapa de anotação dos *markables* em relação à sua anaforicidade (atributos *status*, *is_anaphoric* e *is_bridging*) levou ao seguinte resultado (Tabela 3): das 2377 descrições definidas classificadas, 1428 são da classe *new* (60,05%), confirmando o elevado número de informações novas nos textos. A classe *associative* representa 7,68% do total das classificações, confirmando o baixo número de casos (183) e as referências dêiticas totalizam somente 17 ocorrências (0,69%). A classificação das descrições definidas *old* representa 31,57% do total de casos classificados e foi distribuída na classe *direct* que engloba 407 casos (17,12%), *indirect* com 291 casos (12,24%) e na *encapsultation* apenas 53 casos (2,21%). O resultado final da anotação de correferência irá considerar a anotação de dois sujeitos, comparadas e revisadas pelos próprios anotadores e ainda revisada por um terceiro juiz.

Tabela 2. identificação das configurações morfossintáticas do corpus Summ-it

np_form	**# (%)**	***pro_form***	**# (%)**
def-np	2069 (40,99%)	pes-pro	154 (3,05%)
bare-np	1132 (22,43%)	dem-pro	35 (0,69%)
indef-np	384 (7,60%)	num-ana	27 (0,54%)
def-pn	384 (7,60%)	indef-pro	21 (0,42%)
pn	308 (6,10%)	int-pro	6 (0,12%)
num-np	156 (3,08%)	poss-pro	0 (0%)
quant-np	110 (2,18%)	**Total *pro_form***	243 (4,82%)
coord-np	98 (1,93%)		
dem-np	90 (1,78%)		
poss-np	73 (1,45%)		
int-np	2 (0,04%)		
Total *np_form*	4804 (95,18%)		
Total *markables*:			**5047** (100%)

Tabela 3. Média da anotação de correferência do corpus Summ-it

Classificações		**Média (%)**
status=new		1428 (60,05%)
status=associative		183 (7,68%)
status=deitic		17 (0,69%)
status=old	*is_anaphoric=direct*	407 (17,12%)
	is_anaphoric=indirect	291 (12,24%)
	is_anaphoric=encapsulation	53 (2,21%)
Total de descrições definidas classificadas:		**2377** (100%)

Tabela 4. incidência de relações RST no corpus Summ-it

Relações	# (%)	Relações	# (%)
Elaboration	344 (20,50%)	Contrast	34 (2,02%)
Attribution	170 (10,10%)	Evidence	32 (1,90%)
Parenthetical	166 (9,80%)	Explanation	23 (1,30%)
Same-unit	142 (8,40%)	Means	21 (1,20%)
Interpretation	103 (6,10%)	Non-volitional-cause	18 (1,07%)
Evaluation	82 (4,80%)	Solutionhood	12 (0,70%)
Background	71 (4,20%)	Antithesis	11 (0,60%)
List	70 (4,10%)	Joint	11 (0,60%)
Purpose	68 (4,50%)	Conclusion	8 (0,47%)
Circumstance	60 (3,50%)	Comparison	7 (0,41%)
Concession	56 (3,30%)	Otherwise	4 (0,23%)
Sequence	47 (2,80%)	Restatement	3 (0,17%)
Non-volitional-result	41 (2,40%)	Volitional Cause	2 (0,11%)
Justify	36 (2,10%)	Summary	1 (0,05%)
Condition	34 (2,02%)		
Total de incidências das relações RST:			**1677** (100%)

A Tabela 4 apresenta uma incidência das relações utilizadas para o corpus Summ-it anotadas com a RSTTool. Esses dados são relevantes para a análise textual, a qual permitirá em trabalhos futuros traçar informações valiosas sobre o corpus, como uma tipologia das relações retóricas para o gênero de textos em foco.

5. Considerações Finais

Este artigo apresentou o corpus Summ-it, anotado com informações de discurso visando, prioritariamente, ao estudo da SA. A metodologia de anotação foi apresentada em detalhe e o resultado da anotação realizada foi exemplificado e contabilizado. Apesar das anotações terem sido realizadas de forma independente, o esquema de anotação utilizado possibilita e interrelação entre as anotações. Como trabalho futuro, iremos explorar essa interrelação e suas aplicações na SA. A disponibilidade desse recurso é importante não só para a pesquisa de SA, mas também para explorar outras tarefas de PLN, tais como a resolução automática de correferência e a construção e avaliação de sistemas de PLN que fazem uso da análise de discurso segundo a RST.

Agradecimentos. Este artigo foi realizado com apoio parcial do CNPQ (Processos nº 310488/2005-2 e nº 381063/2005-4) e da Capes/Fulbright.

Referências

Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* PhD thesis, Arhus University.

Carlson, L. and Marcu, D. (2001). Discourse tagging reference manual. Technical Report ISI-TR-545.

Coelho, J. C. B., Collovini, S., and Vieira, R. (2006). Instruções para anotação de relações anafóricas e referência dêitica. Disponível em: http://www.inf.unisinos.br/ renata/laboratorio/guidelines/guidelines-versao2.6.pdf.

Mann, W. C., Matthiessen, C. M. I. M., and Thompson, S. A. (1992). Rhetorical structure theory and text analysis. In Mann, W. C. and Thompson, S. A., editors, *Discourse description: diverse linguistic analyses of a fund-raising text*, Amsterdam. John Benjamins.

Mann, W. C. and Thompson, S. A. (1988). Rhethorical structure theory: toward a functional theory of text organization. *Text*, 8(3).

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

Marcu, D. (1999). Discourse trees are good indicators of importance in text. In Mani, I. and Maybury, M. T., editors, *Advances in Automatic Text Summarization*. The MIT Press.

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

Müller, C., Rapp, S., and Strube, M. (2002). Applying co-training to reference resolution. In *Proc. of the 40th Annual Meeting of the ACL*, Philadelphia, PA.

Müller, C. and Strube, M. (2001). Mmax: A tool for the annotation of multi-modal corpora. In *Proc. of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Seattle, Washington.

Ng, V. and Cardie, C. (2002). Identifying anaphoric and non-anaphoric noun phrases. In *Proc. of the Nineteenth International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan.

O'Donnell, M. (2000). Rsttool 2.4: A markup tool for rhetorical structure theory. In *Proc. of the International Natural Language Generation Conference*, Mitzpe Ramon, Israel.

Ono, K., Sumita, K., and Miike, S. (1994). Abstract generation based on rhetorical structure extraction. In *Proc. of the International Conference on Computational Linguistic - Coling-94*, Japan.

Pardo, T. A. S. (2005). *Métodos para Análise Discursiva Automática*. PhD thesis, ICMC-USP, São Carlos, SP.

Poesio, M. (2004). The mate/gnome scheme for anaphoric annotation, revisited. In Strube, M. and Sidner, C., editors, *Proc. of the 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, Massachusetts, USA.

Poesio, M., Alexandrov-Ksbadjov, M., Vieira, R., Goulart, R., and Uryupina, O. (2005). Does discourse-new detection help definite description resolution? In *Proc. of the 6th International Workshop on Computational Semantics*, Tiburg.

Rino, L. H. M. (1996). *Modelagem de Discurso para o Tratamento da Concisão e Preservação da Idéia Central na Geração de Textos*. PhD thesis, IFSC-USP, São Carlos, SP.