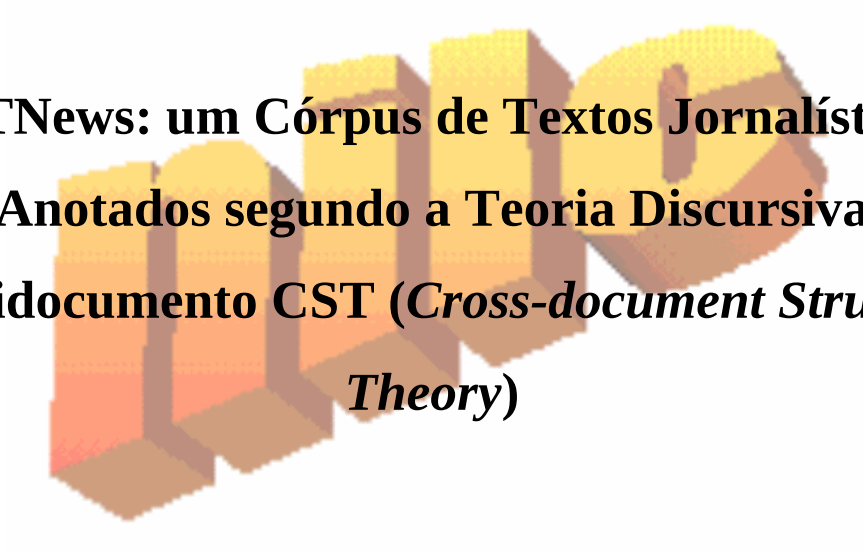
Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP



CSTNews: um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (*Cross-document Structure Theory*)

Priscila Aleixo
Thiago Alexandre Salgueiro Pardo

NILC-TR-08-05

Junho, 2008

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 córpus de textos jornalísticos para o português do Brasil anotados segundo a teoria discursiva multidocumento CST (*Cross-document Structure Theory*). O córpus, chamado CSTNews, é a primeira experiência desse tipo de anotação para o português que se conhece. A construção de tal córpus faz parte do desenvolvimento de um analisador discursivo multidocumento automático para o português, no âmbito de um projeto de Mestrado.

Este trabalho tem o apoio da FAPESP e do CNPq.

Índice

1 Introdução

Apresenta-se, neste relatório, o córpus CSTNews, composto por textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (*Cross-document Structure Theory*) (Radev, 2000). Além disso, este relatório tem como propósito ser um guia para a anotação de córpus nesse estilo.

A CST identifica relações entre sentenças de diferentes documentos que versam sobre assuntos relacionados. Tal informação é útil para diversas tarefas de Processamento de Língua Natural, como sumarização multidocumento, perguntas e respostas e extração e recuperação de informação, por exemplo.

O CSTNews possui atualmente 50 coleções de textos, sendo que cada coleção trata de um assunto diferente e tem em média 4 documentos. Este é o primeiro experimento de identificação de relações CST que faz parte de um projeto de Mestrado que visa à construção do primeiro analisador discursivo multidocumento automático para o português do Brasil.

O relatório está dividido da seguinte forma: a Seção 2 introduz a CST e suas relações; a Seção 3 mostra o processo de anotação do córpus CSTNews e os resultados obtidos; a Seção 4 traz as conclusões e os trabalhos futuros.

2 CST (*Cross-document Structure Theory*)

A CST (*Cross-document Structure Theory*), proposta por Radev (2000), surgiu frente à necessidade da identificação de relações entre vários textos, estruturando o discurso de forma a conectar sentenças provenientes de diferentes documentos e estabelecendo um ou mais tipos de relações entre elas.

As relações podem acontecer entre palavras, sintagmas, orações, sentenças, parágrafos ou blocos de texto ainda maiores. Apesar de orações e sentenças serem tradicionalmente consideradas as unidades discursivas por excelência, tarefas particulares podem exigir um relacionamento entre unidades menores.

Pode-se observar o conjunto de relações na Tabela 1. Essas relações não são as 24 relações definidas no artigo inicial de Radev (2000), mas um refinamento das mesmas proposto por Zhang et at. (2002), que manteve apenas 18 relações. Os nomes das relações foram preservados em inglês, como no trabalho de Zhang. Também são apresentadas as descrições das relações.

Radev afirma que as relações CST não são mutuamente exclusivas, podendo um

mesmo par de segmentos textuais ter mais de uma relação entre si. É interessante notar que, na análise CST, nem todas as sentenças se relacionam.

Tabela 1: Relações CST

Relação	Descrição
Identity	O mesmo texto aparece em mais de um local.
Equivalence (Paraphrase)	Duas sentenças possuem a mesma informação contida.
Translation	Mesma informação, contida em línguas diferentes.
Subsumption	S1 contém toda a informação em S2, mais informação adicional que não está em S2.
Contradiction	S1 e S2 apresentam informação conflitante.
Historical Background	S1 fornecem contexto histórico da informação em S2.
Citation	S1 explicitamente cita o documento S2.
Modality	S1 apresenta uma versão mais qualificada da informação em S2, por exemplo, "é dito que; se sabe que".
Attribution	S1 atribui a versão da informação em S2 usando, por exemplo, "de acordo com a CNN".
Summary	S1 resume S2.
Follow-up	S1 apresenta informação adicional a qual tem acontecido desde S2.
Indirect speech	S1 indiretamente menciona algo o qual foi diretamente mencionado em S2.
Fulfillment	S1 afirma a ocorrência de um evento previsto em S2.
Elaboration (Refinement)	S1 fornece detalhes de alguma informação dada mais generalizada em S2.
Description	S1 descreve uma entidade mencionada em S2.
Reader Profile	S1 e S2 fornecem a mesma informação, porém escrita para ouvintes diferentes.
Change of perpective	A mesma entidade apresenta uma opinião diferente ou apresenta um fato por outro ângulo.
Overlap (partial equivalence)	S1 informa fatos X e Y enquanto S2 informa fatos X e Z; Y e Z devem ser não-triviais.

Na Figura 1, pode-se notar a multiplicidade de relações CST: S1 e S2 podem ser relacionadas pelas relações *Contradiction* e *Attribution.* No primeiro caso, há informações contraditórias: S1 diz que a colisão foi no 26º andar e S2 diz que foi no 25º andar; no segundo caso, a relação *Attribution* se deve ao fato de que a informação contida em S1 e em S2 está sendo atribuída em S1 a uma jornalista, ou seja, a fonte da informação está sendo identificada.

(S1) A colisão no 26o andar ocorreu às 5:50 p.m. na quinta-feira, disse a jornalista Desideria Cavina.

(S2) O avião colidiu no 25o andar do prédio Pirelli no centro de Milão.

Figura 1: Exemplo de identificação de relações CST

É importante ressaltar que na CST algumas relações possuem direcionalidade, por exemplo, as relações *Attribution*, *Subsumption* e *Historical Background*, entre outras. A direcionalidade é mostrada pelos símbolos – (não há direcionalidade), > (direcionalidade de S1 para S2) e < (direcionalidade de S2 para S1).

Veja, por exemplo, duas notícias de diferentes fontes:

(S1) "*Um pequeno avião chocou-se em um edifício no centro de Milão, incendiando os últimos andares do prédio, informou uma jornalista italiana da CNN"*.

(S2) *Um pequeno avião chocou-se hoje com um edifício no centro de Milão incendiando vários andares do prédio.*

Neste par de sentenças duas relações ocorrem: uma de paráfrase ou equivalência e outra de atribuição. Note que na relação de paráfrase não há uma direcionalidade específica, pois tanto S1 é paráfrase de S2, como S2 é paráfrase de S3. Não acontece o mesmo na relação de atribuição, em que a direcionalidade é S1 e S2 (>), pois a atribuição do fato é dada a jornalista em S1 e o mesmo não ocorre se trocarmos a direcionalidade entre S2 e S1 (<).

Veja exemplos com ambas as direcionalidades da relação *Historical Background:*

(S1) "*Este é um dos símbolos financeiros italianos e é um dos prédios mais altos no mundo construído entre 1955 e 1960"*.

(S2) "*Este foi construído de concreto e vidro"*.

Nestas duas sentenças temos como relação *Historical Background* com direcionalidade >, porque é S1 que está trazendo todo o fato histórico do prédio. Já nas duas sentenças abaixo é S2 que traz o fato histórico. Portanto sua direcionalidade é <.

(S1) *O prédio da Pirelli em Milão foi atingido por um avião de pequeno porte.*

(S2) *O prédio foi construído em 1958 e desenhado pelos arquitetos Gio Ponti e Pier Luigi Nervi*.

2.1 Exemplo das relações

Nesta subseção são exemplificadas cada uma das relações definidas em Zhang et al. (2002) e mais a relação nula, isto é, quando não há relação. Os exemplos aqui expostos foram extraídos e traduzidos do CSTBank (Radev et al., 2004), o primeiro córpus em língua inglesa anotado

segundo a CST, também de textos jornalísticos.

1) **Nome da relação**: *Identity*

Descrição: O mesmo texto aparece em mais de um local.

Exemplo: (S1) As vítimas do acidente foram 14 passageiros e três membros da tripulação.

(S2) As vítimas do acidente foram 14 passageiros e três membros da tripulação.

2) **Nome da relação**: *Equivalence/Paraphrase*

Descrição: Duas sentenças possuem a mesma informação contida.

Exemplo: (S1) O avião acidentado levava 14 passageiros e três tripulantes.

(S2) As vítimas do acidente foram 14 passageiros e três membros da tripulação.

3) **Nome da relação**: *Translation*

Descrição: Mesma informação contida em línguas diferentes.

Exemplo: (S1) Gritos de "Viva la revolucion!" ecoaram pela noite.

(S2) Os rebeldes podiam ser ouvidos gritando "Viva a revolução!".

4) **Nome da relação**: *Subsumption*

Descrição: S1 contém toda a informação em S2, mais informação adicional que não está em S2.

Exemplo: (S1) Com três vitórias este ano, o Green Bay tem o melhor record na NFL.

(S2) O Green Bay possui três vitórias este ano.

5) **Nome da relação**: *Contradiction*

Descrição: S1 e S2 apresentam informação conflitante.

Exemplo: (S1) A colisão no 26º andar ocorreu às 5:50 p.m. na quinta-feira, disse a jornalista Desideria Cavina.

(S2) O avião colidiu no 25º andar do prédio Pirelli no centro de Milão.

6) **Nome da relação**: *Historical Background*

Descrição: S1 fornece contexto histórico da informação em S2**.**

Exemplo: (S1) Essa foi a quarta vez que um membro da família real se divorcia.

(S2) Ontem o Duque de Windsor se divorciou da Duquesa de Windsor.

7) **Nome da relação**: *Citation*

Descrição: S1 explicitamente cita o documento S2.

Exemplo: (S1) O príncipe Albert continuou dizendo: "Eu nunca trapaceei".

(S2) Um artigo anterior publicou o príncipe Albert dizendo: "Eu nunca

trapaceei".

8) **Nome da relação**: *Modality*

Descrição: S1 apresenta uma versão mais qualificada da informação em S2, por exemplo, "é dito que; se sabe que".

Exemplo: (S1) Sean "Puffy" Combs é tido como um dos mais ricos.
(S2) Puffy possui milhões de dólares em imóveis na área de Nova York.

9) **Nome da relação**: *Attribution*

Descrição: S1 atribui a versão da informação em S2 usando, por exemplo, "de acordo com a CNN".

Exemplo: (S1) A colisão no 26º andar ocorreu às 5:50 p.m. na quinta-feira, disse a jornalista Desideria Cavina.
(S2) O avião colidiu no 25º andar do prédio Pirelli no centro de Milão.

10) **Nome da relação**: *Summary*

Descrição: S1 resume S2.

Exemplo: (S1) Os Mets ganharam o título em sete jogos.
(S2) Depois dos exaustivos seis primeiros jogos, os Mets ganharam hoje novamente e levaram o título.

11) **Nome da relação**: *Follow-up*

Descrição: S1 apresenta informação adicional a qual tem acontecido desde S2.

Exemplo: (S1) 102 casualidades foram reportadas na área do terremoto.
(S2) Até agora, nenhuma casualidade de abalo foi confirmada.

12) **Nome da relação**: *Indirect Speech*

Descrição: S1 indiretamente menciona algo o qual foi diretamente mencionado em S2.

Exemplo: (S1) O presidente anunciou a população a garantia de novas moradias.
(S2) "Eu garanto novas moradias", disse o presidente à população.

13) **Nome da relação**: *Elaboration/Refinement*

Descrição: S1 informa detalhes de alguma informação dada mais generalizada em S2.

Exemplo: (S1) 50% dos estudantes estão abaixo de 25 anos; 20% estão entre 26 e 30 anos; o restante está acima de 30 anos.
(S2) A maioria dos estudantes da universidade estão abaixo de 30 anos.

14) **Nome da relação**: *Fulfillment*

Descrição: S1 afirma a ocorrência de um evento previsto em S2.

Exemplo: (S1) Após ter viajado para a Áustria quinta-feira, Mr.Green retornou para casa

em Nova York.
(S2) Mr.Green irá para a Áustria quinta-feira.

15) **Nome da relação**: *Description*
Descrição: S1 descreve uma entidade mencionada em S2.
Exemplo: (S1) Várias pessoas ficaram machucadas no prédio da Pirelli após vários andares dos 32 existentes pegarem fogo, um reporte local relatou.
(S2) O prédio da Pirelli possui os escritórios administrativos locais do Lombardy e fica perto da estação central de trem.

16) **Nome da relação**: *Reader Profile*
Descrição: S1 e S2 fornecem a mesma informação, porém escrita para diferentes ouvintes.
Exemplo: (S1) A *durian* é uma fruta usada na cozinha asiática e possui um cheiro forte.
(S2) O prato é geralmente feito com a *durian*.

17) **Nome da relação**: *Change of Perspective*
Descrição: A mesma entidade apresenta uma opinião diferente ou apresenta um fato por outro ângulo**.**
Exemplo: (S1) Oficiais americanos disseram que não havia indicações de que seria um ataque terrorista.
(S2) Anteriormente, em Roma, o presidente do senado, Marcello Pera disse que provavelmente parecia ser um ataque terrorista.

18) **Nome da relação**: *Overlap (partial equivalence)*
Descrição: S1 informa fatos X e Y enquanto S2 informa fatos X e Z; Y e Z devem ser não-triviais.
Exemplo: (S1) Hoje um pequeno avião bateu no 25º andar de um prédio no centro de Milão.
(S2) Um pequeno avião bateu no prédio mais alto do centro de Milão na quinta-feira à noite, expelindo fumaça dos andares mais altos.

19) **Nome da relação**: Não há relação
Descrição: S1 e S2 não possuem nenhum tipo de relação.
Exemplo: (S1) Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.
(S2) As vítimas do acidente foram 14 passageiros e três membros da tripulação.

3 O córpus CSTNews

O córpus construído e anotado, como descrito anteriormente, será a base do estudo para a produção do *parser* (analisador) discursivo automático multidocumento. Como não se tem conhecimento de um córpus para o português do Brasil que satisfaça os requisitos deste trabalho, foi construído o primeiro córpus de notícias jornalísticas anotadas segundo a CST, as quais foram coletadas de fontes distintas.

Esse córpus possui atualmente 50 coleções de textos jornalísticos de domínios variados e cada coleção possui em média 4 documentos de diferentes fontes que versam sobre um mesmo assunto. O número exato de documentos por coleção e assunto, assim como o número de sentenças e palavras por coleção, é mostrado na Tabela 2.

Os textos foram coletados manualmente de jornais on-line por um período de 2 meses, entre Agosto e Setembro de 2007. As fontes dos textos foram os jornais on-line Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo. Essas fontes foram escolhidas devido a grande popularidade na web e também por trazerem as principais notícias do dia corrente, que é o que importa para o córpus, ou seja, uma mesma notícia publicada em fontes diferentes. Os textos jornalísticos foram escolhidos por possuírem uma linguagem clara e do dia a dia, além da facilidade de serem encontrados na web.

Tabela 2: Estatísticas do córpus

Coleção	A	Qt. de documentos	Qt. de sentenças	Qt. de palavras
C1	Mundo	3	24	432
C2	Política	4	78	1405
C3	Cotidiano	4	143	2864
C4	Cotidiano	3	40	846
C5	Cotidiano	2	24	574
C6	Cotidiano	4	50	1253
C7	Ciência	2	23	587
C8	Esportes	3	24	600
C9	Política	4	64	1543
C10	Mundo	5	79	1987
C11	Cotidiano	5	128	2320
C12	Mundo	3	34	974
C13	Mundo	3	37	962
C14	Mundo	4	54	1402
C15	Mundo	4	43	986
C16	Política	3	43	1033
C17	Política	2	49	965
C18	Mundo	3	74	1301
C19	Esportes	2	13	299
C20	Política	5	120	2516
C21	Política	3	41	870
C22	Cotidiano	5	127	2300
C23	Mundo	2	25	572
C24	Esportes	4	52	1091
C25	Esportes	5	159	2788
C26	Mundo	5	116	2621
C27	Esportes	5	181	2985

C28	Esportes	5	70	1336
C29	Mundo	3	48	1167
C30	Dinheiro	3	46	1136
C31	Esportes	2	10	217
C32	Mundo	4	112	2354
C33	Cotidiano	5	131	2803
C34	Cotidiano	3	60	1139
C35	Mundo	5	90	1976
C36	Cotidiano	4	124	2134
C37	Cotidiano	2	27	475
C38	Esportes	5	79	1470
C39	Cotidiano	4	54	1324
C40	Política	5	73	1881
C41	Esportes	5	109	1945
C42	Política	2	40	1074
C43	Política	5	141	1643
C44	Política	2	28	737
C45	Cotidiano	3	50	1226
C46	Mundo	5	78	1519
C47	Mundo	5	99	2753
C48	Esportes	2	43	800
C49	Cotidiano	5	69	575
C50	Política	4	108	2388
Total de documentos		**195**		
Total de sentenças			**3.534**	
Total de palavras				**72.148**

Acredita-se que a quantidade de textos anotados que o córpus possui atualmente será suficiente para o processo extração de conhecimento lingüístico e para a formação de padrões necessários para a construção do *parser* discursivo. Esses mesmos textos serão usados para treino e teste em protótipos que utilizarão aprendizado de máquina para automatizar o processo de identificação das relações.

3.1 O processo de anotação do córpus

O processo de anotação do CSTNews ocorreu durante 3 meses, de Fevereiro à Abril de 2008. Algumas tarefas foram executadas antes do começo da anotação, como seguem: escolha das relações CST, desenvolvimento de uma ferramenta que auxilia o processo de anotação e treinamento dos anotadores.

Após o estudo das relações CST, foi concluído que, para este trabalho, algumas relações definidas por Zhang et al. (2002) seriam redundantes e, portanto, foram excluídas deste trabalho. Das 18 relações definidas por Zhang, 14 delas foram mantidas na anotação do córpus e mais a relação de "Não há relação", quando acontece de um par de sentenças não ser anotado. As relações excluídas foram: *Description*, *Fulfillment*, *Reader Profile* e *Change of Perpective*. Além da exclusão dessas relações, a *Historical Background* que foi definida originalmente como uma relação que traz um fato histórico, neste estudo, é apenas chamada de

Background, retirando assim essa restrição e podendo, então, acontecer entre outros fatos que não sejam apenas históricos.

Como dito, as relações excluídas foram consideradas redundantes frente a outras relações. Por exemplo, a *Description* pode ser uma forma de *Elaboration*. Veja o exemplo a seguir:

(S1) "Várias pessoas ficaram machucadas no prédio da Pirelli após vários andares dos 32 existentes pegarem fogo, um reporte local relatou".

(S2) "O prédio da Pirelli possui os escritórios administrativos locais do Lombardy e fica perto da estação central de trem".

Essas duas sentenças anteriormente tidas como *Description*, são relacionadas por *Elaboration,* pois S2 traz fatos que ajudam a entender melhor uma entidade mencionada em S1.

A relação *Fulfillment* pode ser uma forma de *Follow-up,* porém com direcionalidade diferente. Veja o exemplo a seguir:

(S1) "Após ter viajado para a Áustria quinta-feira, Mr.Green retornou para casa em Nova York".

(S2) "Mr.Green irá para a Áustria quinta-feira".

Como visto na seção 2.1, este exemplo foi classificado como sendo *Fulfillment*, mas mudando a direcionalidade da relação para < (S2 para S1), é possível observar que acontece a relação *Follow up*, sendo assim, a relação *Fulfillment* pode ser contida em *Follow up.*

A relação *Reader Profile* pode ser vista como várias outras. No exemplo a seguir, pode-se usar a relação *Elaboration*, sendo que as sentenças possuem direcionalidade > (S1 para S2).

(S1) A ***durian*** *é uma fruta usada na cozinha asiática e possui um cheiro forte.*

(S2) O prato é geralmente feito com a ***durian****.*

Por fim, a relação *Change of Perpective* é uma forma de *Contradiction.* No exemplo abaixo, duas entidades diferentes contradizem sobre a mesma informação.

(S1) Oficiais americanos disseram que não havia indicações de que seria um ataque terrorista.

(S2) Anteriormente, em Roma, o presidente do senado, Marcello Pera disse que provavelmente

parecia ser um ataque terrorista.

Após o conjunto de relações estar bem definido, foi desenvolvida a ferramenta CSTTool (Aleixo e Pardo, 2008) para facilitar o processo de anotação. A ferramenta oferece um ambiente semi-automático para a anotação CST. Inicialmente, o propósito desta ferramenta é auxiliar os anotadores no trabalho de anotação do córpus, porém, mais adiante, o melhor protótipo de análise discursiva desenvolvido neste trabalho deverá ser acoplado a CSTTool. Ao final de seu desenvolvimento, a CSTTool fornecerá as opções ao usuário de realizar a anotação CST de forma manual, de forma completamente automática ou, ainda, mesclando-se as duas formas nos vários passos que a anotação CST exige. Mais detalhes sobre a CSTTool podem ser encontrados em Aleixo e Pardo (2008).

Com o córpus e a ferramenta para a anotação devidamente prontos, o próximo passo foi o treinamento dos anotadores. Duas pessoas familiarizadas com análise discursiva e com a CST foram treinadas para realizarem a anotação do CSTNews. Um treinamento de 2 horas foi realizado e alguns conjuntos de textos de teste foram anotados para verificar o grau de entendimento e concordância entre os anotadores. Detalhes sobre os resultados da anotação são mostrados a seguir.

3.2 Resultados da Anotação

Observou-se com a anotação que as relações excluídas não fizeram diferença na identificação das mesmas. Além disso, os anotadores não sentiram a necessidade de inclusão de novas relações. De todas as possibilidades de encontrar uma relação entre duas sentenças, 46,83% dessas possibilidades, na verdade, não existiam, pois não havia relação, e 53,17% das possibilidades de fato possuíam algum tipo de relação. A porcentagem de freqüência de cada relação identificada no córpus é exibida na Tabela 3.

Tabela 3: Freqüência das relações no córpus

Relação	Freqüência no córpus
Elaboration	23,98%
Overlap	19,85%
Subsumption	15,24%
Background	6,49%
Atributtion	5,68%
Equivalence	5,09%
Follow-up	4,72%
Contradiction	4,35%
Summary	4,35%
Identity	3,69%
Modality	3,54%
Indirect Speech	2,73%
Citation	0,29%

Translation	0%

Como esperado, algumas relações raramente aparecem, como *Citation e Indirect Speech*, e a relação *Translation* não foi identificada nenhuma vez no córpus.

Para medir a concordância entre os anotadores, foi calculada a medida Kappa (Carletta, 1996) para um conjunto de textos com 3 documentos, 60 sentenças, 1.139 palavras e 1.043 possíveis pares de sentenças relacionadas. Os resultados obtidos da medida Kappa no geral e para as relações identificadas, são mostrados na Tabela 4.

Tabela 4: Concordância entre os anotadores e as relações.

Relação	Kappa
Elaboration	0.321
Overlap	0.562
Subsumption	0.006
Follow-up	0.009
Summary	0.003
Indirect Speech	0.013
Não há relação	0.279
Total	0.258

As relações que obtiveram o maior grau de concordância entre os anotadores, foram *Overlap* (0.562) e *Elaboration* (0.321). A medida Kappa calculada sobre todos os possíveis pares de sentenças foi de 0.258. Segundo Krippendorff (1980), um resultado abaixo de 0.67 é considerado não confiável, porém, há de se considerar que na anotação CST são possíveis 19 rótulos diferentes e não mutuamente exclusivos; o que torna o trabalho muito mais complexo.

Nos experimentos de Radev et al. (2004), a medida Kappa foi calculada em um conjunto de textos com 7.579 possíveis pares de sentenças relacionadas e o resultado obtido foi de 0.4021, não muito diferente do experimento mostrado neste relatório apesar de possuir um número maior de possíveis pares de sentenças relacionadas.

4 Conclusões e Trabalhos Futuros

Neste relatório, foi apresentada a primeira experiência de anotação de um córpus multidocumento para o português do Brasil, o CSTNews.

As contribuições desta anotação foram: o refinamento da teoria CST, que foi de grande valia para tornar a anotação mais fácil, e também a construção de uma ferramenta semi-automática para a anotação, a CSTTool.

Com esses primeiros resultados, o projeto passa para uma nova etapa, a identificação de possíveis padrões para cada uma das relações e posteriormente a tentativa de automatizar todo o processo, acoplando o analisador automático a CSTTool. Prevê-se, também, a extensão do córpus com mais textos anotados.

Referências

Aleixo, P. e Pardo, T.A.S. (2008). *CSTTool: Uma ferramenta semi-automática para anotação de córpus pela teoria discursiva multidocumento CST*. Série de Relatórios do NILC. NILC-TR-08-03. São Carlos-SP, Maio, 11p.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. 22(2):249–254.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In *the Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*. Hong Kong.

Radev, D.R.; Otterbacher, J.; Zhang, Z. (2004). CST Bank: A Corpus for the Study of Cross-document Structural Relationships. In the *Proceedings of Fourth International Conference on Language Resources and Evaluation*.

Zhang, Z.; Blair-Goldensohn, S.; Radev, D.R. (2002). Towards CST-enhanced summarization. In the *Proceedings of the AAAI 2002 Conference*. Edmonton, Alberta.