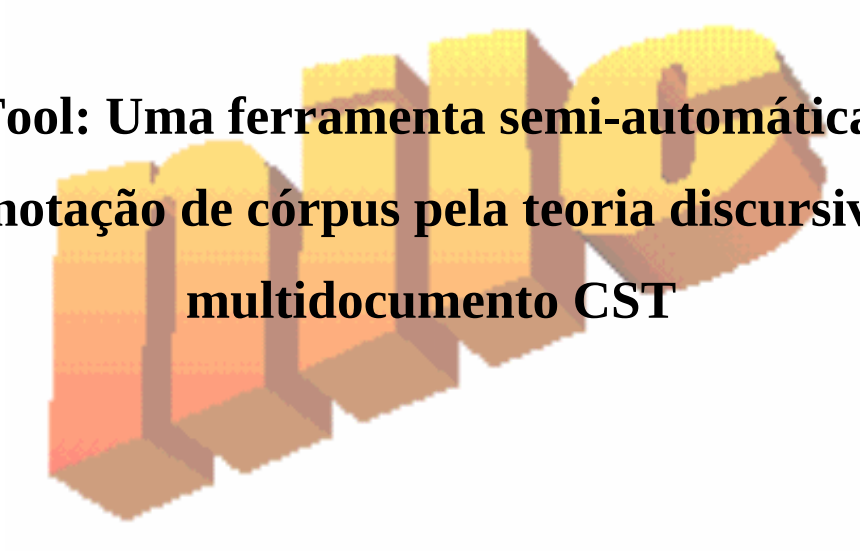
Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP



CSTTool: Uma ferramenta semi-automática para anotação de córpus pela teoria discursiva multidocumento CST

Priscila Aleixo

Thiago Alexandre Salgueiro Pardo

NILC-TR-08-03

Maio, 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 descreve a construção de uma ferramenta de anotação de córpus chamada CSTTool. A ferramenta auxilia o usuário a analisar textos via a teoria discursiva multidocumento CST (*Cross-document Structure Theory*). A construção de tal ferramenta 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 1

2 CST (*CROSS-DOCUMENT STRUCTURE THEORY*) 1

3 A FERRAMENTA CSTTOOL 4

4 CONCLUSÕES E TRABALHOS FUTUROS 9

REFERÊNCIAS BIBLIOGRÁFICAS 9

APÊNDICE A 10

1 Introdução

Neste relatório, são relatadas as funcionalidades da ferramenta CSTTool desenvolvida para anotação de córpus pela teoria discursiva multidocumento CST (*Cross-document Stucture Theory*) proposta pro Radev (2000).

Diante da necessidade de se anotar um córpus de textos jornalísticos, o objetivo da construção da CSTTool é proporcionar um ambiente semi-automático para facilitar a anotação. O córpus faz parte do projeto de desenvolvimento do primeiro *parser* discursivo automático multidocumento para o Português do Brasil.

As principais funcionalidades da ferramenta neste primeiro estágio são: segmentação sentencial automática ou manual, detecção de segmentos possivelmente relacionados e marcação das relações encontradas entre os segmentos. Pretende-se, mais adiante, acoplar à CSTTool a automatização do último processo citado acima.

A próxima seção apresenta uma breve revisão bibliográfica da teoria CST. A seguir, na Seção 3, relatam-se as funcionalidades da ferramenta desenvolvida e, por fim, na Seção 4, conclusões e 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. Baseada na RST (*Rhetorical Structure Theory*) (Mann e Thompson, 1987), a CST visa auxiliar em tarefas de processamento da língua de natureza multidocumento, como perguntas e respostas e sumarização automática multidocumento, dentre outras.

Radev, em seu trabalho original, define 24 relações discursivas para relacionamento intertextual, que pode se dar 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.

Em um trabalho mais elaborado, Zhang et al. (2002) refinam essas relações, produzindo um conjunto de 18 relações, as quais podem ser vistas na Tabela 1. Os nomes das relações foram preservados em inglês, como no trabalho original. Também são apresentadas as descrições das relações. S1 e S2 referem-se a segmentos textuais (sentenças, normalmente) provenientes de fontes diferentes.

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.

A Figura 1 ilustra duas sentenças relacionadas. Ambas foram publicadas em fontes diferentes, porém trazem a mesma informação, a de que haviam 14 passageiros e 3 tripulantes em um acidente de avião. Não se pode afirmar que essas duas sentenças são idênticas, mas se pode dizer que são equivalentes, ou seja, transmitem a mesma mensagem.

Fonte X

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

Fonte Y

O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes.

Figura 1: Exemplos de sentenças de diferentes fontes.

Radev afirma que as relações CST não são mutuamente exclusivas, podendo um mesmo par de segmentos textuais ter mais de uma relação. É interessante notar que, na análise CST, nem todas as sentenças se relacionam e o importante é o valor da sentença, o que ela explicita, para, assim, haver algum tipo de relacionamento.

Na Figura 2, pode-se notar a multiplicidade de relações CST: as sentenças 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 2: 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. Na relação de paráfrase não há uma direcionalidade específica, pois tanto S1 é paráfrase de S2, quanto S2 é paráfrase de S3. Não acontece o mesmo na relação de atribuição, em que a direcionalidade é de S1 para S2 (>), pois a jornalista em S1 é a fonte do fato descrito em S2.

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.*

3 A Ferramenta CSTTool

Como etapa inicial para o desenvolvimento do primeiro *parser* discursivo multidocumento automático, tem-se a anotação manual de um córpus de textos jornalísticos relacionados. Diante da custosa anotação manual e da explosão de possibilidades dessa anotação, foi proposta a CSTTool, que oferece um ambiente semi-automático para a anotação CST. Inicialmente, o propósito desta ferramenta é auxiliar os anotadores no trabalho de anotação do córpus. A arquitetura atual da ferramenta é ilustrada na Figura 3.

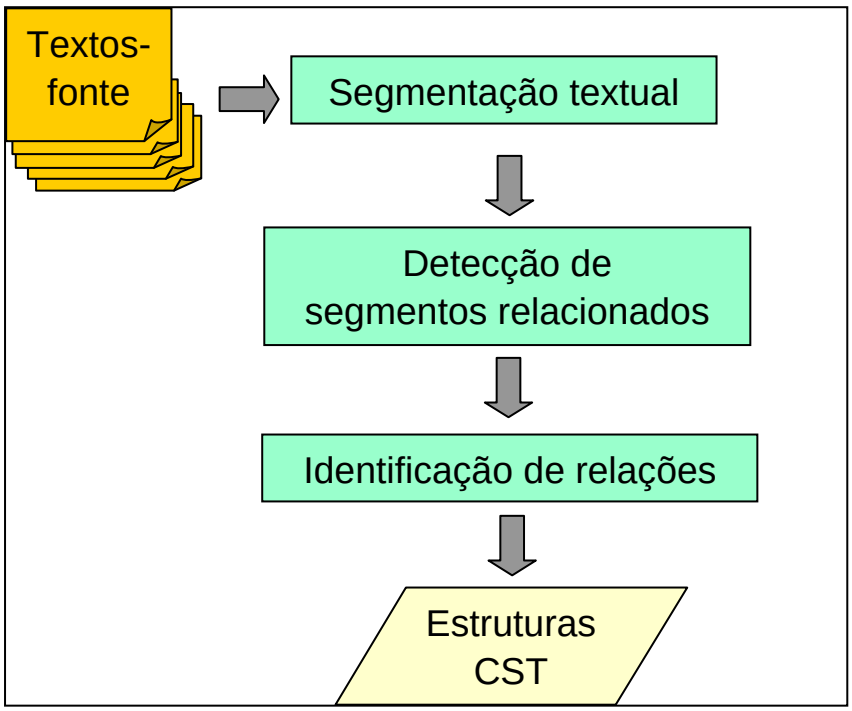


Figura 3: Arquitetura da CSTTool.

O primeiro passo para a anotação de dois textos relacionados é segmentá-los em sentenças. Nessa parte, o anotador pode escolher fazer a segmentação automática ou manual. A segmentação automática é realizada pelo SENTER (Pardo, 2006), o segmentador sentencial automático desenvolvido no NILC. A Figura 4 ilustra a primeira tela da ferramenta referente à etapa de segmentação sentencial.

A tela intitulada *Text Segmentation* possibilita segmentar um texto automaticamente ou manualmente, como já mencionado. Para a segmentação automática, que se encontra na parte superior da tela, abre-se o arquivo texto desejado em *Open* e aciona-se o botão *Segmente File*. Opcionalmente, após a segmentação, abre-se o arquivo gerado *.seg* e, clicando em *Open Text*, o texto segmentado aparecerá na parte inferior da tela. Alterações poderão ser feitas se desejadas e o anotador poderá salvá-las, clicando em *Save segmented text.*

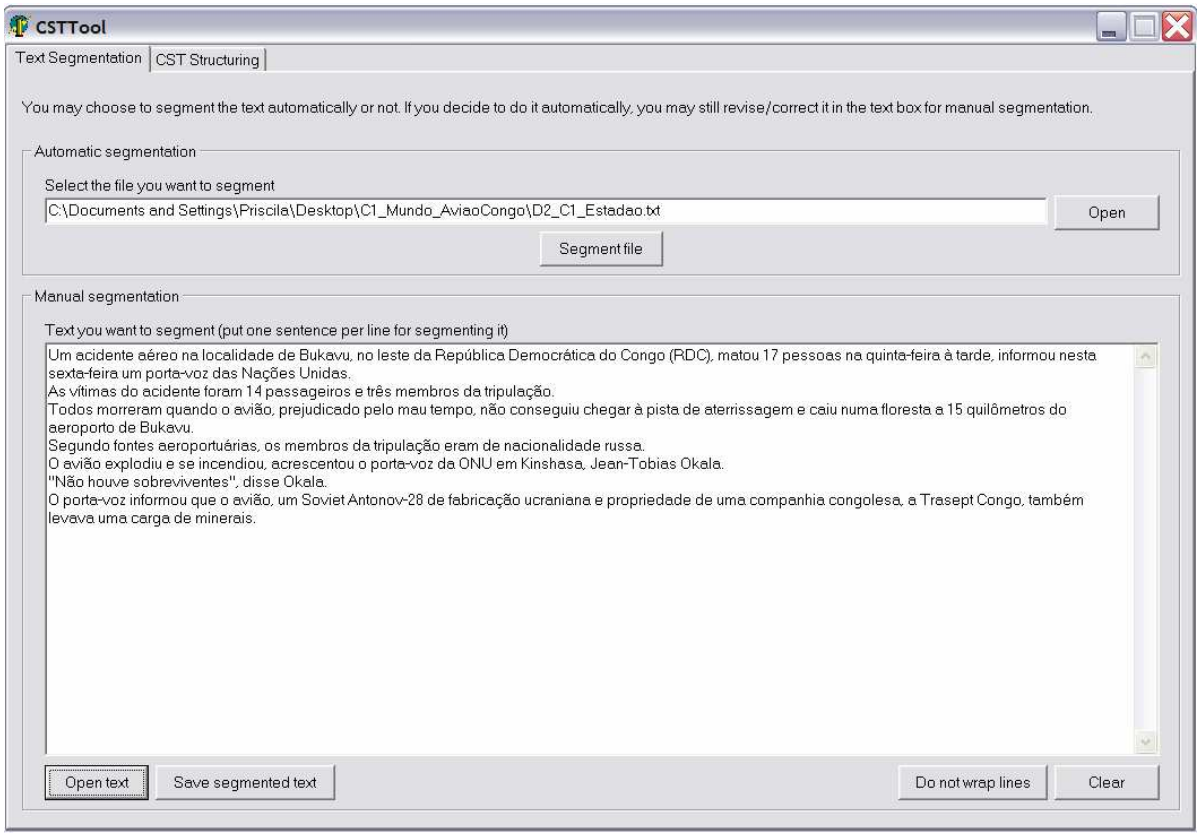


Figura 4: Tela de segmentação da CSTTool.

Para a segmentação manual, basta clicar em *Open Text*, na parte inferior da tela e segmentar o texto escolhido manualmente, pulando linha a cada mudança de sentença. Os mesmos passos devem ser repetidos para salvar o arquivo segmentado como descrito acima. Mais um recurso possível nessa tela é o de *Do not wrap lines*, em que há a opção de visualizar o texto com ou sem quebra de linha.

Na segunda tela da ferramenta, como ilustra a Figura 5, encontra-se a etapa de estruturação dos textos via a teoria CST. Nesta tela, poderá ser realizada a anotação semi-automática de textos, identificando relações entre pares de sentenças provenientes de diferentes textos. No estado atual da ferramenta, somente dois textos poderão ser anotados ao mesmo tempo. Esta delimitação foi escolhida devido à complexidade de se identificar as relações em mais de 2 textos simultaneamente.

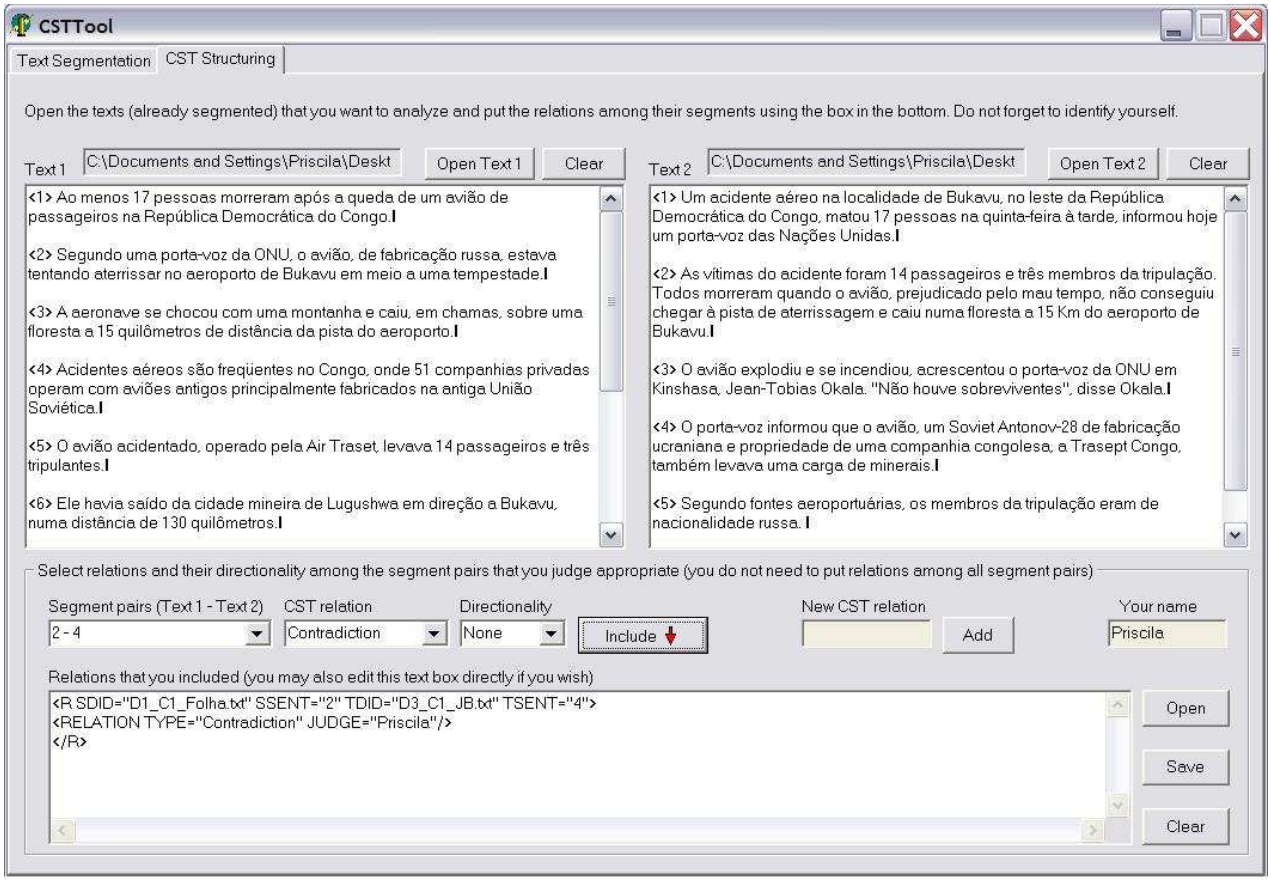


Figura 5: Tela da estruturação CST na CSTTool.

Nesta tela, primeiramente, o anotador deve identificar-se no campo *Your name*. Em seguida, clicando em *Open Text1*, procura-se por um texto já segmentado e salvo no formato *.seg*. O mesmo passo é feito para *Open Text2*. Com os dois textos visíveis na tela, clicando-se na caixa *Segment pairs*, é possível verificar quais pares de sentenças são fortes candidatas a possuírem uma relação CST. Esse resultado se dá pelo cálculo automático da medida *WordOverlap*, dispensando assim, o trabalho árduo de se ter que olhar todas as combinações de sentenças dos 2 textos.

Como descrita por Radev et al. (2008), a medida *WordOverlap* é dada por:

$$Wol(S1,S2) = \frac{\#CommonWords}{\#Words(S1) + \#Words(S2)}$$

em que se divide o número de palavras em comum entre as duas sentenças pela soma do número de palavras nas duas sentenças.

A medida *WordOverlap* se faz necessária devido ao grande espaço de busca que deverá ser explorado, sendo inviável a análise sentença a sentença manualmente, o que daria um número exponencial de comparações. Segundo Radev et al. (2008), de todas as medidas de

similaridade por eles testadas, essa foi a que se mostrou mais eficiente.

Depois de detectado de fato que duas sentenças possuem uma relação, deve-se escolher o tipo de relação na caixa *CST relation,* sua direcionalidade em *Directionality* e clicar no botão *Include*. Em princípio, a CSTTool vem com as 18 relações previstas por Zhang et al. (2000). Para adicionar uma nova relação, deve-se escrever seu nome na caixa de texto intitulada *New CST relation* e clicar em *Add*.

Caso haja mais relações possíveis em um mesmo par de sentenças ou direcionalidades diferentes, é só repetir o processo acima. Toda a identificação de relação será mostrada na parte inferior da tela em formato XML.

O formato XML é muito usado em anotação de córpus, pois utiliza marcações para descrever, ou mesmo classificar, o tipo de informação de um texto. Para este trabalho, optou-se em usar este formato não só pela sua grande utilização, mas também pelo fato de Radev et al. (2004) terem utilizado o formato em seu córpus multidocumento, o CST Bank. Na CSTTool, segue-se o mesmo formato de codificação deste córpus.

Tomando como exemplo a saída XML mostrada na Figura 5, tem-se a seguinte informação:

- Etiquetas <R> e </R> - começo e fim de uma relação;
- Atributo SDID – nome do documento;
- Atributo SSENT – número da sentença correspondente ao documento;
- Atributo TDID – nome do outro documento;
- Atributo TSENT – número da sentença correspondente ao outro documento;
- Atributo RELATION TYPE – tipo de relação atribuída ao par de sentenças;
- Atributo JUDGE – nome do anotador.

A direcionalidade na anotação é dada pela ordem em que os nomes dos documentos aparecem. No exemplo da Figura 5, há uma relação de Contradição entre a sentença 2 do documento 1 e a sentença 4 do documento 3. Neste caso, apesar da ordem dos documentos serem 1 e 3, também é verdade afirmar que acontece uma relação de Contradição entre as sentenças dos documentos 3 e 1, porém não é necessário replicar a anotação, pois se sabe que a relação de Contradição não possui direcionalidade.

Uma análise completa de um par de textos pode ser encontrada no Apêndice A.

4 Conclusões e Trabalhos Futuros

Neste relatório, foi apresentada a ferramenta CSTTool para anotação semi-automática de córpus segundo a teoria CST. Essa ferramenta é de suma importância para o desenvolvimento do primeiro *parser* discursivo multidocumento automático para o português do Brasil, pois os textos anotados na ferramenta nortearão a construção do parser.

Além de a ferramenta facilitar o processo de anotação manual das relações CST, esta também poderá futuramente hospedar o *parser* em desenvolvimento, proporcionado ao usuário a possibilidade de fazer uma anotação semi-automática ou totalmente automática.

Referências Bibliográficas

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

Pardo, T.A.S. (2006). *SENTER: Um Segmentador Sentencial Automático para o Português do Brasil*. Série de Relatórios do NILC. NILC-TR-06-01. São Carlos-SP, Janeiro, 6p.

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*.

Radev, D., Otterbacher, J., Zhang, Z. (2008). Cross-document Relationship Classification for Text Summarization. In: *Computational Linguistics* (to appear).

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

Apêndice A

A seguir, é apresentada a anotação completa de um par de textos realizada na CSTool. O Texto 1 é ilustrado na Figura 6 e o Texto 2 na Figura 7. A anotação XML produzida pela CSTTool é exibida na Figura 8.

<1> O presidente do Conselho de Ética do Senado, Leomar Quintanilha (PMDB-TO), disse hoje ser contrário à unificação dos processos contra o senador Renan Calheiros (PMDB-AL) que tramitam na Casa Legislativa.
<2> A unificação foi proposta nesta terça-feira pela bancada do PT no Senado.
<3> Na opinião de Quintanilha, uma única investigação sobre diversas denúncias não permite conclusões distintas sobre o suposto envolvimento de Renan em ações que possam configurar a quebra do decoro parlamentar.
<4> "Eu defendo a tese de que cada representação deve ter tratamento distinto. Vamos unificar se eu for voto vencido", disse.
<5> Quintanilha afirmou que vai manter a "coerência" em sua decisão, uma vez que negou pedido do PSOL para incluir as novas denúncias contra Renan no primeiro processo contra o peemedebista que entrou na pauta do plenário do Senado na semana passada.
<6> "Eu vou manter a coerência da minha decisão. Eu acho que vão confundir as acusações se tudo for analisado conjuntamente. Eu advogo a causa do tratamento distinto a cada representação."
<7> A unificação dos processos tem o apoio da bancada do PT no Senado e do PSOL, partido que ingressou com três das quatro representações contra Renan.
<8> O senador José Nery (PSOL-PA) questionou apenas o fato da unificação estar sendo discutida neste momento, já que há alguns meses o Conselho de Ética negou o pedido da legenda para reunir as denúncias contra Renan em uma única peça.
<9> "Tentamos fazer um aditamento, o conselho nos respondeu que não seria possível. O nosso entendimento é que todas as representações tratam de possível quebra de decoro. Esse conjunto de denúncias poderia ser agregado em um único processo. Ficou fatiado nesse monte de processos porque o conselho assim o quis", criticou Nery.
<10> Quintanilha disse que, mesmo se a unificação dos processos for aprovada pelo conselho, vai defender relatores distintos para cada uma das denúncias contra Renan.
<11> Mas reconheceu que, dificilmente, os relatores terão posições semelhantes nas representações.
<12> "Quem garante que vão trabalhar ao mesmo tempo, com cronograma similar? E se um tiver um entendimento sobre as denúncias e os outros não concordarem? Isso dificulta os trabalhos", afirmou.
<13> Na opinião do senador Aloizio Mercadante (PT-SP), o plenário da Casa não deve analisar separadamente as três representações contra Renan que tramitam no Senado.
<14> "Nós devemos analisar as três representações que faltam, oferecer um relatório completo das três para que cada um forme definitivamente o seu julgamento de mérito", defendeu.
<15> O Conselho de Ética vai decidir sobre a unificação dos processos contra Renan em reunião marcada para a próxima quarta-feira.
<16> No mesmo dia, o conselho vai discutir o relatório do senador João Pedro (PT-AM) sobre o segundo processo contra o peemedebista --no qual é acusado de usar sua influência política para beneficiar a empresa Schincariol.
<17> O conselho decidiu deixar a reunião para a semana que vem uma vez que, amanhã, a Mesa Diretora do Senado vai decidir se encaminhará a quarta representação contra Renan ao Conselho de Ética da Casa.
<18> Quintanilha achou melhor esperar a decisão da Mesa antes de discutir a unificação dos processos, já que a quarta representação poderá se reunir às outras duas contra o peemedebista caso todos os processos passem a tramitar juntos no Conselho de Ética.

Figura 6: Texto 1.

<1> O Conselho de Ética do Senado vai decidir na próxima semana sobre o pedido do líder do P-SOL, José Nery (PA), para juntar em uma só as representações apresentadas contra o presidente da Casa, Renan Calheiros (PMDB-AL).
<2> O presidente do conselho, Leomar Quintanilha (PMDB-TO), disse que é contra a união das representações, mas que vai colocar a proposta em votação.
<3> - Ainda que o objeto seja o mesmo - a perda de mandato - cada representação tem uma natureza.
<4> Sou contra porque a junção pode confundir os assuntos - afirmou.
<5> Para o senador José Nery, a união das representações "demonstra mais robustez e elementos de convencimento para mostrar, na nossa avaliação, que houve quebra de decoro parlamentar ".
<6> Ele disse que a exemplo do que ocorreu com a primeira representação, o ideal é que seja criada uma comissão de três relatores para analisar os outros processos em conjunto.
<7> Na quinta-feira, a Mesa Diretora do Senado se reúne às 14 horas para decidir se aceita a quarta representação contra o presidente da Casa.
<8> A representação pede a investigação de denúncia de que Renan coordenaria um esquema para arrecadar recursos junto a ministérios comandados pelo PMDB.

Figura 7: Texto 2.

```
<R SDID="D1_C44_Folha.txt.seg" SSENT="1" TDID="D2_C44_JB.txt.seg" TSENT="1">
<RELATION TYPE="Overlap" JUDGE="Priscila"/>
</R>
<R SDID="D1_C44_Folha.txt.seg" SSENT="1" TDID="D2_C44_JB.txt.seg" TSENT="2">
<RELATION TYPE="Overlap" JUDGE="Priscila"/>
</R>
<R SDID="D1_C44_Folha.txt.seg" SSENT="3" TDID="D2_C44_JB.txt.seg" TSENT="2">
<RELATION TYPE="Elaboration" JUDGE="Priscila"/>
</R>
<R SDID="D2_C44_JB.txt.seg" SSENT="2" TDID="D1_C44_Folha.txt.seg" TSENT="4">
<RELATION TYPE="Indirect-speech" JUDGE="Priscila"/>
</R>
<R SDID="D2_C44_JB.txt.seg" SSENT="3" TDID="D1_C44_Folha.txt.seg" TSENT="4">
<RELATION TYPE="Elaboration" JUDGE="Priscila"/>
</R>
<R SDID="D1_C44_Folha.txt.seg" SSENT="5" TDID="D2_C44_JB.txt.seg" TSENT="1">
<RELATION TYPE="Follow-up" JUDGE="Priscila"/>
</R>
<R SDID="D1_C44_Folha.txt.seg" SSENT="5" TDID="D2_C44_JB.txt.seg" TSENT="2">
<RELATION TYPE="Follow-up" JUDGE="Priscila"/>
</R>
<R SDID="D2_C44_JB.txt.seg" SSENT="2" TDID="D1_C44_Folha.txt.seg" TSENT="6">
<RELATION TYPE="Indirect-speech" JUDGE="Priscila"/>
</R>
<R SDID="D1_C44_Folha.txt.seg" SSENT="6" TDID="D2_C44_JB.txt.seg" TSENT="2">
<RELATION TYPE="Elaboration" JUDGE="Priscila"/>
</R>
<R SDID="D1_C44_Folha.txt.seg" SSENT="9" TDID="D2_C44_JB.txt.seg" TSENT="1">
<RELATION TYPE="Follow-up" JUDGE="Priscila"/>
</R>
<R SDID="D2_C44_JB.txt.seg" SSENT="1" TDID="D1_C44_Folha.txt.seg"
TSENT="15">
<RELATION TYPE="Subsumption" JUDGE="Priscila"/>
</R>
<R SDID="D1_C44_Folha.txt.seg" SSENT="17" TDID="D2_C44_JB.txt.seg"
TSENT="7">
<RELATION TYPE="Subsumption" JUDGE="Priscila"/>
</R>
```

Figura 8: Saída XML da CSTTool.