# UNIVERSIDADE DE SÃO PAULO

## Instituto de Ciências Matemáticas e de Computação

**ISSN 0103-2569**


---

**MANUAL DE ANOTAÇÃO DE SINALIZADORES DISCURSIVOS EM TEXTOS JORNALÍSTICOS**


EWERSON DANTAS
LARISSA JESUS SANTA BÁRBARA
MATEUS ARAÚJO PEREIRA
NAIRA SILVA GAMA
TOBIAS JOSÉ ARAÚJO ALMEIDA
JACKSON WILKE DA CRUZ SOUZA
PAULA CHRISTINA FIGUEIRA CARDOSO
ROANA RODRIGUES



**Nº 447**


---

# RELATÓRIOS TÉCNICOS


São Carlos – SP
**Ago./2024**

*Natural Language Processing initiative (NLP2) of the Center for Artificial Intelligence (C4AI) of the University of São Paulo, sponsored by IBM and FAPESP*

# Manual de Anotação de Sinalizadores Discursivos em Textos Jornalísticos

EWERSON DANTAS
LARISSA JESUS SANTA BÁRBARA
MATEUS ARAÚJO PEREIRA
NAIRA SILVA GAMA
TOBIAS JOSÉ ARAÚJO ALMEIDA
JACKSON WILKE DA CRUZ SOUZA
PAULA CHRISTINA FIGUEIRA CARDOSO
ROANA RODRIGUES

Agosto/2024

**Relatório Técnico do**
**Núcleo Interinstitucional de Linguística Computacional (NILC)**

# SUMÁRIO

# 1. Introdução

Este manual foi elaborado para a anotação de sinalizadores das relações retóricas do modelo teórico RST (*Rhetorical Structure Theory*) (Mann e Thompson, 1988). A ideia principal da teoria é a de que um texto coerente é formado por unidades mínimas de discurso (*Elementary Discourse Units* - EDUs ou proposições) que desempenham funções retóricas para que o objetivo comunicacional do autor, por meio da realização discursiva no texto seja atingido.

Carlson e Marcu (2001) caracterizam unidades mínimas de discurso como orações. Tais unidades são ligadas umas às outras por meio de relações retóricas (também chamadas de relações de coerência ou discursivas), formando uma estrutura discursiva conectada, geralmente representada na forma de árvore. Segundo Taboada e Mann (2006), inicialmente a RST objetivava o desenvolvimento de um modelo que pudesse ajudar na automação da geração de textos, no entanto, foi adotada por pesquisadores de diferentes áreas e para diversas finalidades, como ensino, descrição e processamento de linguagem natural (PLN), auxiliando na melhor compreensão do texto e na proposta de uma estrutura conceitual das relações de coerência.

Embora a teoria RST exista há pelo menos três décadas, sua relevância para o PLN permanece significativa até os dias atuais. Um estudo recente conduzido por Souza, Cardoso e Rodrigues (2023) exemplifica essa relevância ao analisar uma lista abrangente de trabalhos sobre RST publicados na última década. A análise revela não apenas a contínua aplicação da teoria em diversas áreas do PLN, mas também seu papel crucial na compreensão e na geração de textos, destacando-se como uma ferramenta fundamental para estruturar e interpretar informações textuais de maneira coerente e eficaz.

As relações retóricas são comumente determinadas com base nos marcadores discursivos (MDs) (ou "conectivos") presentes em um texto. Tais MDs são tidos como elementos que estabelecem relações entre proposições, papel desempenhado de maneira majoritária na gramática por meio de conjunções e preposições. A literatura (Marcu; 2000, Pardo; 2005, Taboada e Das; 2013) indica que a identificação de marcadores em função da relação a que ocorrem facilita o processamento do texto. Porém, a ausência de um MD prototípico não furta a possibilidade de interpretação entre proposições.

Estudos recentes (Das e Taboada; 2018, Liu e Zeldes; 2019) apontam que as relações RST podem ser identificadas a partir de sinais associados às relações RST que vão além dos MDs. Por não marcarem a relação e não serem exclusivos, parece que a noção de Sinalizadores Discursivos (SDs) é mais adequada, quando comparada a de MDs nesse caso.

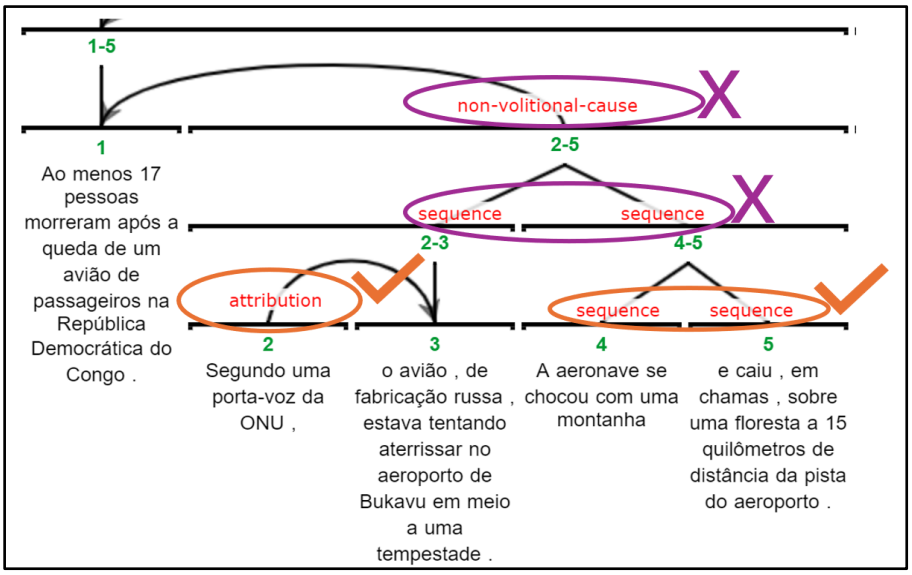
Assim, **nosso objetivo** aqui é demonstrar possibilidades de ocorrência dos sinalizadores de relações RST em textos jornalísticos. Neste relatório, portanto, descrevemos o manual de anotação de sinalizadores discursivos no *corpus* CSTNews (Cardoso *et al.*, 2011), o qual contém textos jornalísticos anotados com RST. O *corpus* está estruturado em 50 conjuntos (*clusters*), com dois ou três textos que noticiam o mesmo evento. Por essa característica multidocumento e de redundância, a anotação deve ser feita apenas no maior texto do conjunto, pois acreditamos que quanto maior for o texto, maior é a chance de encontrarmos mais relações RST e, possivelmente, essas relações ocorram nos outros outros textos da mesma coleção.

A partir de um estudo piloto (Rodrigues; Souza; Cardoso, 2023), foi feita a anotação em uma pequena porção de textos do *corpus* CSTNews. Os resultados demonstram uma série de pistas linguísticas e estruturais que potencialmente indicam as relações RST, mas ainda não haviam sido catalogadas em língua portuguesa. A partir disso, pretendemos avançar nos estudos para apresentar uma descrição do *corpus* e propor uma taxonomia de organização e compreensão de SDs em função das relações RST.

Este manual foi construído refletindo a metodologia de anotação de Rodrigues, Souza e Cardoso (2023) com relação à natureza da ocorrência dos sinalizadores nas sentenças. Assim como os autores, anotaremos apenas sinalizadores intrassentenciais.

Na Figura 1, as relações RST estão ocorrendo entre diferentes EDU's que ocorrem em diferentes segmentos discursivos. Conforme destacado, as relações destacadas com ***X*** não seriam anotadas pois estão acontecendo entre duas sentenças (portanto, tem um comportamento intersentencial). Já as relações destacadas com ✓devem ser anotadas, já que ocorrem dentro do mesmo segmento sentencial, tendo um comportamento *intrasentencial*.

**Figura 1** - Exemplo de anotação RST




Fonte: *Corpus* CSTNews (Cardoso *et al.*, 2011).

Além de indicarmos o esquema de anotação dos sinalizadores, também mostramos como realizar a instalação da ferramenta rstWeb (Zeldes, 2016), que será utilizada nesta tarefa. Ademais, com base no estudo preliminar de Rodrigues, Souza e Cardoso (2023), apresentamos a anotação de algumas relações RST em função dos sinalizadores já catalogados na taxonomia apresentada neste manual.

Este relatório técnico é produto do projeto de pesquisa "Relações retóricas para além de marcadores discursivos: explorando a anotação RST para o Português Brasileiro", que visa caracterizar sinalizadores linguístico-estruturais de relações RST. O projeto faz parte de um projeto maior, intitulado POeTiSA (*POrtuguese processing - Towards Syntactic Analysis and parsing*) que objetiva aumentar os recursos baseados em sintaxe e discurso e desenvolver ferramentas e aplicações relacionadas para a língua portuguesa brasileira, buscando alcançar resultados mundiais de última geração nesta área.

Este relatório está organizado em duas seções, além desta introdução. Na Seção 2, descrevemos o que se sabe sobre os SDs identificados até agora. Na Seção 3, é detalhado o processo de anotação propriamente dito, assim como a ferramenta utilizada para realizar as anotações.

# 2. O que sabemos até agora?

Identificar relações RST por meio de marcas explícitas no texto não é uma tarefa nova, especialmente em PLN em tarefas de análise de discurso. Os MDs são as pistas na superfície do texto que mais são investigadas na literatura. Isso se deve ao fato de, em muitos casos, eles serem capazes de identificar, articular e caracterizar relações de coerência entre proposições (Da Cunha *et al.*; 2012, Hernault *et al.*; 2010, Marcu; 2000, Pardo; Nunes; 2008, Liu e Zeldes; 2019).

Nesta seção, descreve-se o *corpus* de trabalho, a taxonomia de sinalizadores discursivos que foram catalogados, o processo de anotação e exemplos de SDs em relações RST.

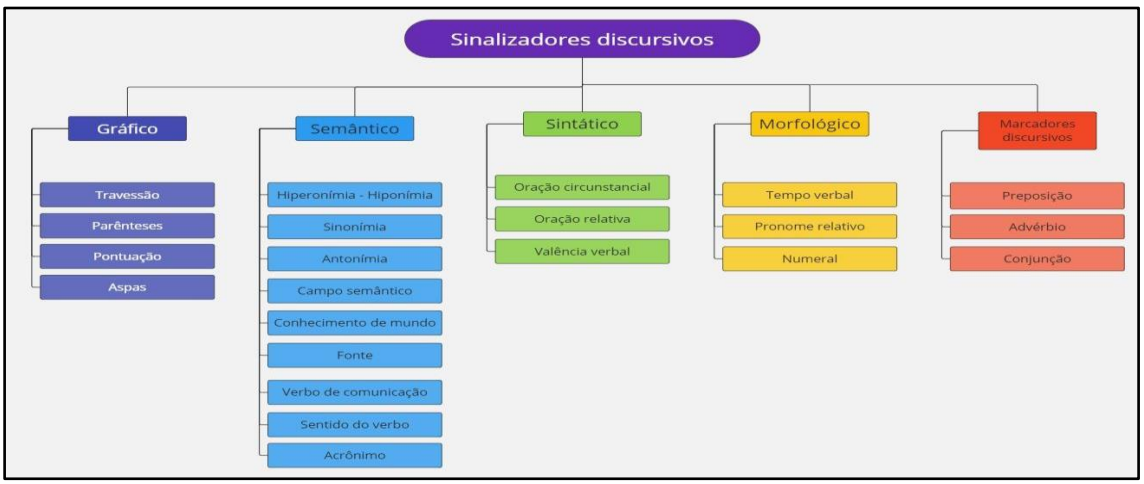
## 2.1 Sobre o *corpus* CSTNews

O *corpus* CSTNews possui 50 conjuntos de textos, organizados por assunto, que foram coletados manualmente no ano de 2007 (Cardoso et al., 2011). No total, são 140 textos, que juntos contabilizam 2,088 sentenças e 47,240 palavras. As fontes dos textos foram os jornais on-line Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo. O CSTNews foi eleito pelo fato de que está anotado com RST.

## 2.2 Sobre a taxonomia

Como dito, a proposta da taxonomia é o aprofundamento de um estudo piloto realizado em *corpus*. Na Figura 2, apresentamos o estado atual da taxonomia. Para sugerir uma classificação adequada ao encontrar novos sinalizadores ou propor novas categorias, é essencial compreender as categorias que já existem na taxonomia.

**Figura 2** - Taxonomia de Sinalizadores Discursivos.



**a) Sinalizadores do tipo marcador discursivo**

Em nossa proposta, esta categoria congrega elementos linguísticos provenientes de diferentes classes gramaticais. Como dito anteriormente, os MDs podem caracterizar prototipicamente determinadas relações RST, como "se" para *Condition*, e "mas" para *Contrast*.

Entretanto, a simples ocorrência de uma preposição, por exemplo, em um segmento discursivo não o faz ser, de fato, um MD de uma relação RST. Nesse sentido, Taboada e Das (2013), utilizando a proposta de Fraser (2009), propuseram condições para que determinado item lexical fosse considerado um MD. De acordo com a proposta, um marcador sinaliza uma relação binária sobre uma única sequência discursiva, composta por trechos adjacentes a uma relação RST (2a); além disso, o MD pode estar posicionado no começo, no meio ou no final da sentença (2a e 2b); MDs não criam relações entre parágrafos e sentenças, mas apenas de maneira intrassentencial (2c e 2d); por fim, MDs apenas orientam a interpretação da relação, sem haver uma relação de exclusividade.

(1)

a) ✓ Lula oscila de 48% para 50%, *enquanto* Alckmin cai de 39% para 36%.

b) ✓ *Segundo* fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.

c) ✓ O tiroteio é um dos piores crimes do tipo no campus de uma universidade nos EUA *desde que* Charles Whitman abriu fogo.

d) **X** A pesquisa de hoje apresentou uma variação na lista espontânea (...). *Neste cenário*, Lula sobe de 27% para 31% (...).

Os MDs podem ser classificados em três subtipos, conforme apresentado no Quadro 1.

**Quadro 1 -** Exemplos de MDs como SDs

| SUBTIPO | EXEMPLO |
|---|---|
| **Preposição** <PREP> | *Segundo* fontes *aeroportuárias, os membros da tripulação eram de nacionalidade russa* ***Relação: Attribution \| Documento: D3_C1]*** |
| **Conjunção** <CONJ> | (...) havia 110 km de congestionamento em toda a cidade *enquanto* a média para o horário era de 76 km. ***[Relação: Comparison \| Documento: D1_C4]*** |
| **Advérbio** <ADV> | *Até o momento*, as autoridades do Sri Lanka não confirmaram as mortes ou esclareceram o que acontece na cidade de Muttur. ***[Relação: Circumstance \| Documento: D1_C13]*** |

**b) Sinalizadores do tipo morfológico**

Essa categoria comporta sinalizadores que indicam morfologicamente informações temporais, ou ainda determinadas classes de palavras que auxiliam na organização discursiva. Diante disso, propusemos três subtipos de sinalizadores, descritos e ilustrados no Quadro 2.

**Quadro 2** - Exemplos de sinalizadores discursivos morfológicos

| SUBTIPO | EXEMPLO |
|---|---|
| **Tempo verbal** <TEMP-VERBAL> | *"Na sexta-feira choveu 12 centímetros em algumas regiões, e há previsão de mais tempestades hoje"* ***[Relação: Sequence \| Documento: D1_C23]*** |
| **Pronome relativo** <PRO-REL> | A temperatura deve permanecer baixa, por conta da massa de ar polar que acompanha a frente fria *que* passa pelo estado. ***[Relação: Elaboration \| Documento: D3_C4]*** |
| **Numeral** <NUM> | A *primeira* iniciou às 10h, e a *segunda* está marcada para às 16h. ***[Relação: List \| Documento: D5_C20]*** |

O *Tempo verbal* pode indicar sucessão de eventos, como no exemplo, devendo, nesse caso, ser considerada a relação entre pretérito e presente entre os dois verbos (a saber, "choveu" e "há"). Todos os verbos envolvidos na sinalização do tempo verbal devem ser considerados e anotados juntos com a mesma etiqueta.

O sinalizador *Pronome relativo* foi considerado um sinalizador em nossa tipologia e que pode, eventualmente, indicar relações de explicação (*Explanation*) ou de detalhe (*Elaboration)* no modelo RST.

Por fim, o sinalizador *Numeral* indica, de alguma forma, a ordenação da informação no texto, como a relação entre "primeira" e "segunda", e "10h" e "16h", indicando uma lista de eventos.

**c) Sinalizadores do tipo sintático**

Nessa classe de sinalizadores enquadramos construções e usos sintáticos que auxiliam na indicação da relação RST, bem como na organização da informação no texto. Em nossa anotação, os sinalizadores sintáticos podem ser divididos nos subtipos apresentados no Quadro 3.

**Quadro 3** - Exemplos de sinalizadores discursivos sintáticos

| SUBTIPO | EXEMPLO |
|---|---|
| **Oração circunstancial** <ORA-CIRCUNS> | (...) desde que *foi solto*, Abadia fez quatro cirurgias plásticas (...) ***[Relação: Sequence \| Documento: D5_C35]*** |

| | |
|---|---|
| **Oração relativa** <ORA-REL> | A temperatura deve permanecer baixa, por conta da massa de ar polar que acompanha a frente fria que *passa pelo estado*. ***Relação: Elaboration \| Documento: D3_C4]*** |
| **Valência verbal** <VALE-VERB> | Já *o rio Tâmisa*, que está com seu leito no limite, *pode transbordar* durante a próxima madrugada. ***[Relação: Same unit \| Documento: D2_C23]*** |

O sinalizador *Oração circunstancial* comporta orações adverbiais (como causais, comparativas, concessivas, condicionais, etc), como "desde que foi solto", que impõem determinadas circunstâncias para que o trecho em seguida seja interpretado. O MD não deve ser anotado em Oração circunstancial, já que ele será anotado em outra categoria.

O sinalizador *Oração relativa* comporta orações adjetivas (reduzidas ou não) explicativas e restritivas, como "que passa pelo estado". Quando acompanhar pronome relativo este não deve ser anotado em Oração relativa, já que ele será anotado em outra categoria.

A *Valência verbal* foi pensada para os casos em que houvesse concordância verbal ou com o sujeito ou com o complemento, como em "o Rio Tâmisa (...) pode transbordar…", conforme o exemplo demonstrado. Tanto sujeito e verbo quanto verbo e complemento devem ser anotados juntos com a mesma etiqueta.

### d) Sinalizadores do tipo semântico

Nessa classe de sinalizadores são consideradas relações semânticas que se estabelecem entre *tokens* do texto e que, potencialmente, podem indicar relações RST. Em nossa anotação, os sinalizadores semânticos podem ser classificados nos subtipos apresentados no Quadro 4.

**Quadro 4** - Exemplos de sinalizadores discursivos semânticos

| SUBTIPO | EXEMPLO |
|---|---|
| **Hiperonímia e Hiponímia** <HIPER-HIPO> | *O degelo da neve também influi no aumento do nível dos rios. O rio Severn, o maior do país, está cinco metros acima do nível normal de verão.* ***[Relação: Explanation \| Documento: D1_C23]*** |
| **Sinonímia** <SINO> | *Este foi o sétimo triunfo consecutivo dos brasileiros na competição --antes, o país conquistou quatro vitórias contra a seleção argentina e duas diante de Portugal.* ***[Relação: Parenthetical \| Documento: D2_C8]*** |
| **Antonímia** <ANTO> | *No segundo turno, as intenções de voto do presidente Lula caíram de 53% em junho para 50% em julho, enquanto o candidato Alckmin subiu de 29% para 36%.* ***[Relação: Contrast \| Documento: D2_C2]*** |
| **Campo semântico** <CAMPO-SEM> | *No começo da rebelião quatro pessoas ficaram feridas, entre elas uma auxiliar de enfermagem e um agente de polícia que trabalham no presídio.*" ***[Relação: Elaboration \| Documento: D2_C37]*** |

| | |
|---|---|
| **Conhecimento de mundo** <CONHE-MUND> | *Um total de 549 pessoas morreram, 3043 ficaram feridas e 295 estão desaparecidas em razão das enchentes.*<br>***[Relação: Non-volitional cause \| Documento: D1_C12]*** |
| **Fonte** <FONTE> | *"É* um *par de irmãos admirável, cada um com cerca de 1% da massa do Sol", disse Jayawardhana."*<br>***[Relação: Attribution\| Documento: D1_C7]*** |
| **Verbo de comunicação** <VERB-COMUNI> | "É um par de irmãos admirável, cada um com cerca de 1% da massa do Sol", *disse* Jayawardhana."<br>***[Relação: Attribution\| Documento: D1_C7]*** |
| **Sentido verbal** <SENT-VERBAL> | "*A chuva começou na noite de domingo e ficou mais forte entre 6h e 7h30 desta segunda.*"<br>***[Relação: Sequence \| Documento: D3_C4]*** |
| **Acrônimo** <ACRO> | As informações coletadas pela PF durante as investigações foram enviadas ao *TJ (Tribunal de Justiça) do Estado de Rondônia* e ao *STJ (Superior Tribunal de Justiça)*.<br>**[Relação: Parenthetical \| Documento: D1_C9]** |

A relação *Hiperonímia-Hiponímia* sinaliza relações RST por meio da utilização de informações mais genéricas e específicas, como "rios" e "rio Severn", respectivamente demonstrados no exemplo. Tanto hiperônimo quanto hipônimo devem ser anotados juntos e com a mesma etiqueta.

O sinalizador *Sinonímia* indica relações por meio de dois ou mais tokens de mesmo sentido, como "triunfo" e "vitórias". Neste sinalizador os tokens podem apresentar flexões distintas de gênero e/ou número. Os tokens que forem identificados como sinônimos devem ser anotados juntos e com a mesma etiqueta.

O sinalizador *Antonímia* indica a relação RST por meio de dois ou mais tokens de sentidos opostos, como "caíram" e "subiu". Neste sinalizador os tokens podem apresentar flexões distintas de gênero e/ou número. Os tokens que forem identificados como antônimos devem ser anotados juntos e com a mesma etiqueta.

*Campo semântico* é o sinalizador que se estabelece entre dois ou mais tokens que congregam o mesmo campo de ideias, que podem ou não fazer parte do mesmo paradigma flexional, como "rebelião" e "presídio". Todos os tokens envolvidos (dois ou mais) no campo semântico devem ser anotados juntos e com a mesma etiqueta.

Já *Conhecimento de mundo* sinaliza as relações RST por meio do acesso ao conhecimento externo ao texto, como saber que "enchentes" é uma possível causa não intencional para mortes. Todos os tokens envolvidos (dois ou mais) no conhecimento de mundo devem ser anotados juntos e com a mesma etiqueta.

O sinalizador *Fonte* identifica a autoria do conteúdo reportado, como "*Jayawardhana*". Tanto determinantes que precedem a Fonte (como "O jornalista disse…") quanto detalhes pós-

postos a ela (como “João, o jornalista, disse…) devem ser anotados, com exceção da pontuação. Todos os tokens envolvidos devem ser anotados juntos e com a mesma etiqueta.

*Verbo de comunicação* é o sinalizador que identifica construções verbais ligadas à forma como o conteúdo informacional é reportado, como “disse”. Devem ser anotados verbos plenos (como “disse”) e/ou locuções verbais (como “tinha dito). No caso das locuções, todos os tokens devem ser anotados juntos e com a mesma etiqueta.

Já o *Sentido verbal* indica semanticamente alguma relação RST, podendo, nesse caso, ser expresso por apenas um verbo, como “começou” que indica o início do evento relatado. Devem ser considerados na anotação verbos plenos ou locuções verbais.

Por fim, *Acrônimo* indicam a explicação de um nome ou ainda a aglutinação desse nome por meio de suas iniciais ou siglas, como “TJ (Tribunal de Justiça”). Devem ser anotados todos os tokens que compõem o acrônimo e/ou sigla, ignorando-se os aspectos gráficos (como parênteses e vírgulas).

**e) Sinalizadores do tipo gráfico**

Os sinalizadores dessa classe caracterizam-se por marcar informações extras, ou explicações, ou siglas acerca de algum tópico no texto. Em nossa anotação, os sinalizadores gráficos se dividem em quatro subtipos, descritos no Quadro 5.

**Quadro 5** - Exemplos de sinalizadores discursivos semânticos

| SUBTIPO | EXEMPLO |
|---|---|
| **Travessão** <TRAVES> | A falha no reversor - mecanismo que ajuda o avião a frear - (...)” **[Relação: Parenthetical \| Documento: D1_C3]** |
| **Parênteses** <PARENT> | "...porque esta é a primeira CNI/Ibope com a lista oficial dos candidatos do TSE ( Tribunal Superior Eleitoral ) **[Relação: Parenthetical \| Documento: D1_C2]** |
| **Pontuação** <PONT> | Uma das três vagas será ocupada pelo major-brigadeiro Allemander Jesus Pereira Filho, indicado para exercer o cargo em substituição a Jorge Luiz Brito Velozo…” **[Relação: Explanation \| Documento: D1_C5]** |
| **Aspas** <ASPAS> | “É um par de irmãos admirável, cada um com cerca de 1% da massa do Sol", *disse* Jayawardhana. **[Relação: Attribution \| Documento: D1_C7]** |

O sinalizador *Travessão* sinaliza a introdução de falas ou diálogos no texto, ou ainda o isolamento ou separação de alguma informação complementar, como demonstrado no exemplo. Devem ser anotados todos os travessões que ocorrerem na mesma sentença com a mesma etiqueta.

Já os *Parênteses* sinalizam o isolamento ou a separação de informações complementares. Devem ser anotados todos os parênteses que ocorrerem na mesma sentença com a mesma etiqueta.

A *Pontuação* é um sinalizador que comporta todos os usos de pontuações que isolam ou separam informações complementares e/ou explicativas.

Por fim, as *Aspas* sinalizam o discurso direto ou o deslocamento de sentidos. As aspas (simples ou duplas) devem ser anotadas juntas e com a mesma TAG.

## 2.3 Sobre o processo de anotação e observação dos sinalizadores

### a) Anotação RST no *corpus*

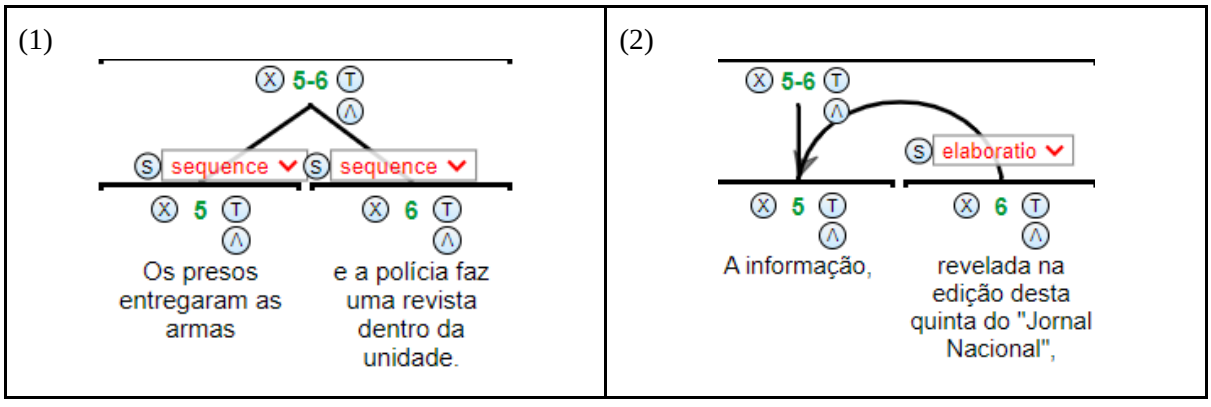
Nesta tarefa de anotação, estamos utilizando o CSTNews, pré-anotado com o modelo teórico RST. Dada concordância entre os anotadores do *corpus* à época, a anotação é considerada válida, não devendo ser mudada.

Caso o anotador identifique algum possível equívoco na anotação, deve sinalizá-lo no formulário e submetê-lo à discussão na reunião ordinária do grupo. Para mais informações sobre a anotação RST, veja Cardoso et al. (2011).

### b) Sinalizador em função da relação RST

É necessário ter atenção ao indicar um sinalizador, pois ele deve estar atrelado à relação RST observada. Nesse caso, a simples ocorrência do sinalizador no texto pode não dizer que a relação, de fato, ocorre como nos casos apresentados na **Figura 3**.

**Figura 3 -** Exemplos de anotação de sinalizadores RST




Na Figura 3, em (1), apesar de os tokens “presos”, “polícia”, “revista” e “unidade” fazerem parte de um mesmo campo semântico, parecem não sinalizar a relação *Sequence* entre as duas EDUs. Já em (2), as aspas no trecho “Jornal Nacional” não sinalizam a relação *Elaboration*, e, por conta disso, nesse caso, não devem ser anotadas. Por conta disso,

destacamos a importância de considerar a relação RST observada, procurando indicar quais sinalizadores potencialmente a indicam.

**c) Situações de atenção**

As situações listadas a seguir exigem atenção ao processo de anotação:

- As **locuções**, de maneira geral, são entendidas como sequências de palavras que possuem o mesmo valor gramatical, como "por enquanto", "até o momento" ou "vai acontecer". Ao indicar uma locução como sinalizador, todo o segmento deverá ser anotado;
- A ferramenta de anotação rstWeb[1] **não permite selecionar partes de palavras**. Sendo assim, ao identificar uma característica morfológica toda a palavra deverá ser indicada;
- Desconsidere todo e qualquer tipo de sinalizador que indique relação RST entre dois ou mais **segmentos sentenciais**;
- Em casos de **relações multinucleares** (como *List* e *Same unit*), identifique os sinalizadores que ocorrem, registrando a anotação no núcleo mais à esquerda;
- Caso identifique algum sinalizador discursivo e não conseguir identificar alguma palavra/*token* correspondente a ele, **em último caso**, indique o tipo e o subtipo do sinalizador na EDU correspondente na relação RST no formulário.

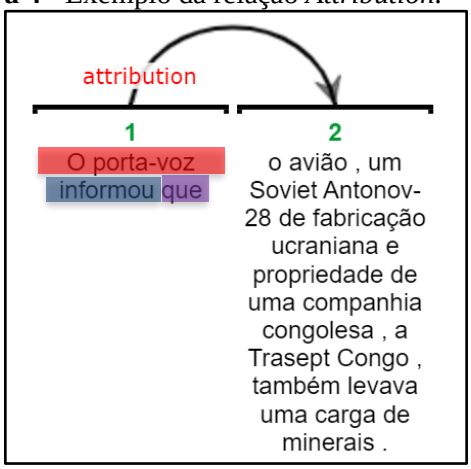
## 2.4 Sobre as relações RST

Nesta seção apresentamos alguns SDs identificados em estudos preliminares. Destacamos que não esgotamos as possibilidades de sinalização das relações, pois trata-se apenas de exemplos prototípicos. Nesse sentido, pode haver outras pistas que não foram ainda identificadas, bem como relações RST ainda não vistas durante o processo de estudo. Ademais, estão exemplificadas apenas as relações que ocorreram durante a elaboração deste manual.

a) *Attribution*: Essa relação significa fala, pensamento ou transmissão de conteúdo por meio do discurso direto ou indireto no texto. No *corpus* analisado, é uma das ocorrências mais frequentes, visto que os textos que o compõem são jornalísticos. Na **Figura 4**, apresenta-se um exemplo em que esta relação é indicada por Fonte de informação, Verbo de comunicação e Conjunção .

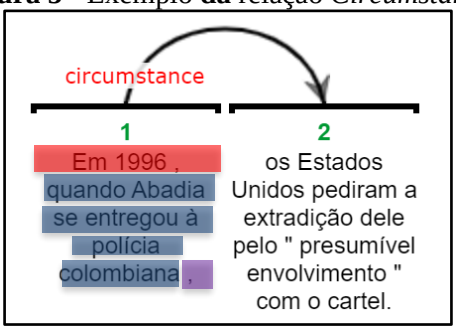
[1] A ferramenta está detalhada na Seção 3.

**Figura 4** - Exemplo da relação *Attribution*.



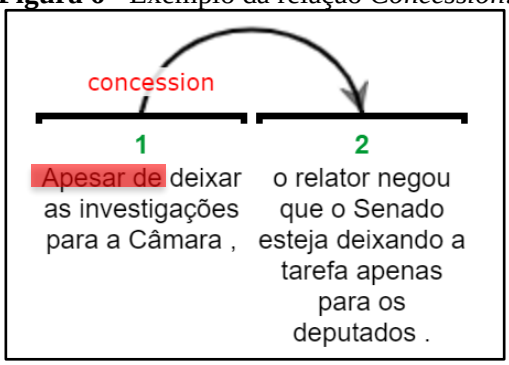

b) *Circumstance:* Essa relação RST deve apresentar uma situação realizável, em que o satélite (EDU 1, na Figura 5 provê a situação que é apresentada no núcleo (EDU 2, na Figura 5). No exemplo abaixo, tem-se que essa relação está sendo sinalizada por meio de Advérbio, Oração circunstancial e Pontuação.

**Figura 5** - Exemplo **da** relação *Circumstance*.




c) *Concession*: Essa relação evidencia uma proposição concessiva, em que o autor organiza *retoricamente* seu discurso apontando que há compatibilidade entre o núcleo (EDU 2, na Figura 6) e o satélite (EDU 1, na Figura 6), porém sem comprometer-se que o satélite pode ser válido. No exemplo a seguir, foi identificado apenas a Locução conjuntiva como sinalizador desta relação.

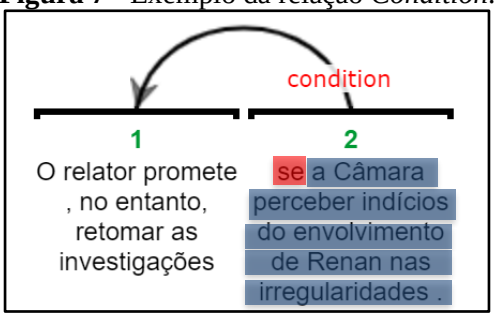
**Figura 6** - Exemplo da relação *Concession*.




d) *Condition*: Essa relação RST comporta condições, em que o satélite (EDU 2, na Figura 7) apresenta uma situação possível para que a situação descrita no núcleo (EDU 1, na

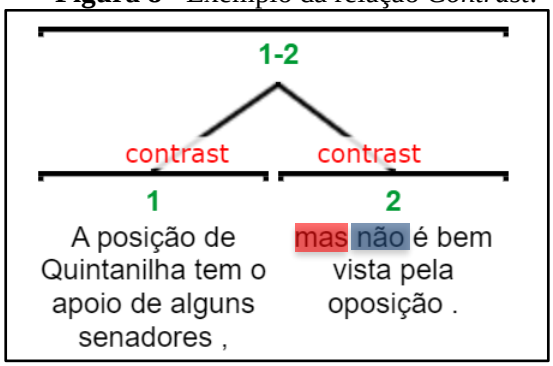
Figura 7) seja realizada. No exemplo a seguir, foram apontados Conjunção e Oração circunstancial como sinalizadores dessa relação.

**Figura 7** - Exemplo da relação *Condition*.



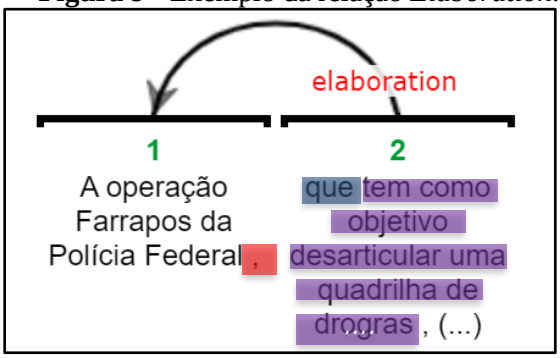

e) *Contrast*: Essa relação é multinuclear e binária, ou seja, sempre acontece entre dois núcleos (como as EDUs 1 e 2, na Figura 8), em que podem ser compreendidas em alguns *aspectos*, bem como serem diferentes e comparadas em relação às diferenças. No exemplo, a seguir, foram identificados dois sinalizadores do tipo MD, a saber Conjunção e Advérbio. **Por ser uma relação multinuclear, os sinalizadores devem ser indicados na EDU sempre mais à esquerda.**

**Figura 8** - Exemplo da relação *Contrast*.



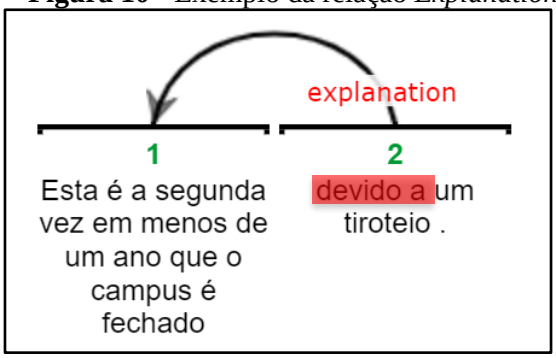

f) *Elaboration*: Essa relação RST é abundante no *corpus* CSTNews. Em seu satélite (como na EDU 2, da Figura 9) apresentam-se detalhes sobre alguma situação ou elemento *presente* no núcleo (como na EDU 1, da Figura 9). No exemplo a seguir, a relação *Elaboration* foi sinalizada por meio da Pontuação, Pronome relativo e Oração relativa que ocorreram no texto original, em Português.

**Figura 9** - Exemplo da relação *Elaboration*.



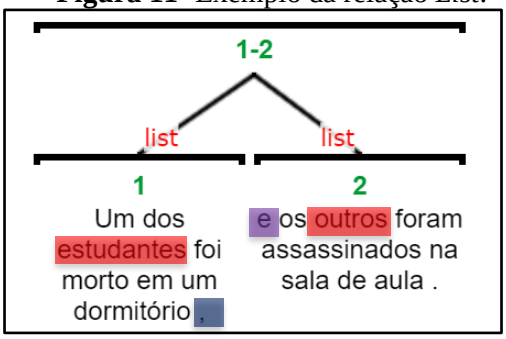

g) *Explanation*: É uma relação que tem como foco a explicação apresentada no satélite (como na EDU 2, na Figura 10) do motivo ou a maneira de como o evento ou *situação* principal tenha acontecido no núcleo (como na EDU 1, na Figura 10). No exemplo analisado, o MD Preposição sinaliza essa relação.

**Figura 10** - Exemplo da relação *Explanation*.



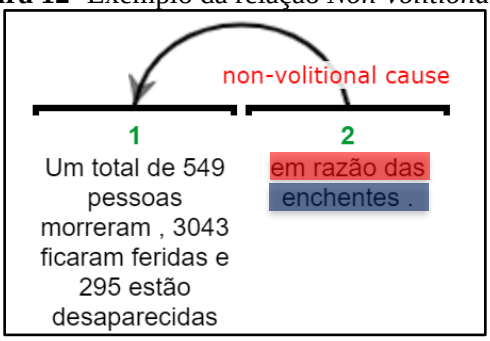

h) *List*: Essa relação RST é multinuclear, fazendo com que seus núcleos (como nas EDUs 1 e 2, *da* Figura 11) apresentem itens que sejam comparáveis, sem nenhuma hierarquia entre si. No exemplo analisado, os sinalizadores Hiperonímia-Hiponímia, Pontuação e Conjunção foram identificados. **Por ser uma relação multinuclear, os sinalizadores devem ser indicados na EDU sempre mais à esquerda.**

**Figura 11**- Exemplo da relação *List*.



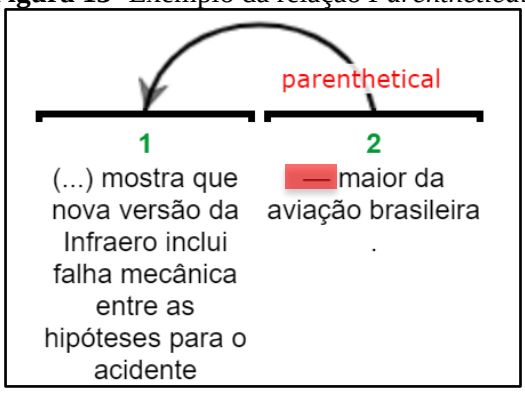
i) *Non-volitional cause*: Essa relação tem a característica de apresentar a causa não intencional no satélite (como na EDU 2, da Figura 12) em relação à situação principal *apresentada* no núcleo (como na EDU 1, na Figura 12). No exemplo analisado, foram identificados a Preposição e o Conhecimento de mundo como sinalizadores dessa relação.

**Figura 12**- Exemplo da relação *Non-volitional cause*.



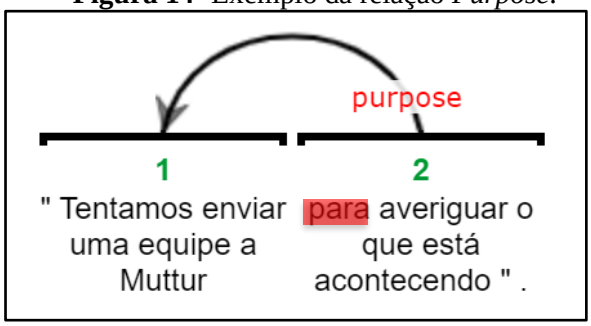

j) *Parenthetical*: Nessa relação RST o satélite (como na EDU 2, na Figura 13) é responsável por apresentar uma informação adicional relacionada ao núcleo (como na EDU 1, na Figura 13). Essa informação extra não faz parte da composição sintática do texto, tendo algum delimitador gráfico para isso. Nesse sentido, o sinalizador Travessão foi identificado no exemplo analisado abaixo.

**Figura 13**- Exemplo da relação *Parenthetical*.



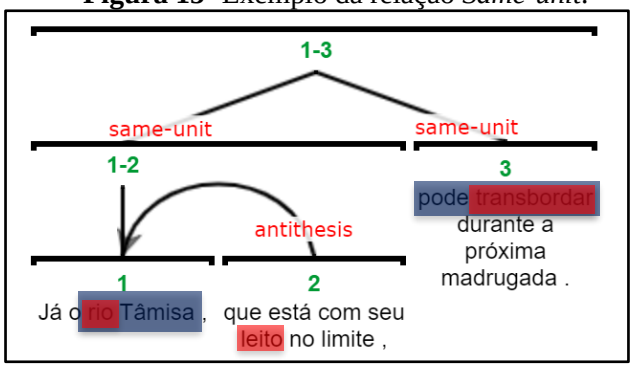

k) *Purpose*: Essa relação RST é caracterizada por ter uma ação apresentada em seu núcleo (como na EDU 1, na Figura 14) e em seu satélite (como na EDU 2, na Figura 14) *uma* situação que pode realizar o núcleo. No exemplo, a Preposição foi identificada como sinalizador dessa relação.

**Figura 14**- Exemplo da relação *Purpose*.



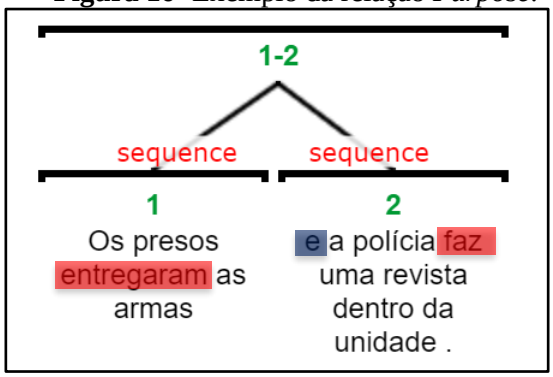

l) *Same-unit*: Essa relação RST também é multinuclear, em que seus núcleos (EDUs 1 e 3, na Figura 15) juntos constituem uma mesma proposição que está segmentada, na maioria das *vezes*, por outra relação RST (como a EDU 2, na Figura 15). No exemplo analisado, os sinalizadores identificados foram Campo semântico e Valência verbal.

**Figura 15**- Exemplo da relação *Same-unit*.




m) *Sequence*: Nessa relação multinuclear os núcleos (como as EDUs 1 e 2, na Figura 16) apresentam eventos que são realizados sequencialmente. No exemplo analisado a seguir, foram identificados os sinalizadores Tempo-Verbal e Conjunção para essa relação. **Por ser uma relação multinuclear, os sinalizadores devem ser indicados na EDU sempre mais à esquerda.**

**Figura 16**- Exemplo da relação *Purpose*.



## 2.5 Como resolver as dúvidas?

Durante a anotação, é possível que surjam dúvidas quanto à escolha de uma etiqueta ou sinalizador para dada relação RST. Para minimizar esse impacto, sugerimos os seguintes procedimentos:

- Leia atentamente o **manual de anotação**, nele estão as definições e catalogações dos sinalizadores e etiquetas;
- Considere a leitura do **material bibliográfico** que serviu de elaboração para o manual de anotação;
- **Tente ao máximo associar** etiquetas e classificações existentes ao sinalizador identificado;
- Associe a etiqueta <CPD> (“casos para discussão”) ao sinalizador que você ficar em dúvida para ser discutido em reunião;
- **Não proponha nenhuma nova ou classificações** (de tipo ou subtipo) sem que haja aprovação da coordenação;
- Não ultrapasse o tempo de **1:30 de anotação diária**;
- **Registre seu trabalho** ao terminar de anotar cada texto;
- Ao identificar **possíveis desvios de anotação RST ou segmentação das sentenças**, registre no formulário e converse com seu orientador;
- **Entre em contato** com os membros de sua subequipe de anotação diante da persistência de dúvidas a fim de minimizar CPD. Se ainda houver dúvidas, consulte os coordenadores do projeto (E-mails: jacksoncruz@ufba.br | paulastm@gmail.com | roana@academico.ufs.br).

# 3. Instalação e configuração da ferramenta rstWeb

Nesta seção, mostramos como instalar a ferramenta de anotação rstWeb, bem como sua configuração. Como material suplementar, sugere-se assistir ao vídeo "rstWeb: instalação e uso[2]".

## 3.1 Instalação da ferramenta no Windows (local) - rstWeb

1. **Baixar o ambiente Anaconda -** Para baixar, acesse [https://www.anaconda.com/download]
2. **Baixar a ferramenta rstWeb -** Para baixar a ferramenta, acesse [https://gucorpling.org/rstweb/info/] e descompacte o programa
3. **Instalar Anaconda -** Para instalar o ambiente, veja este tutorial [Instalando o Jupyter - Pacote Anaconda para Programação em Python]
4. **Abrir a ferramenta Jupyter -** No Jupyter, acesse a pasta onde está o programa e crie um novo notebook de código
5. **Comandos python -** No novo notebook, insira os seguintes comandos em linhas diferentes: (i) "pip install cherrypy"; (ii) "pip install selenium" e (iii) "!rstweb_local.bat"
6. **Acessar rstWeb -** Acessar a ferramenta rstWeb no seu browser utilizando o seguinte link [http://127.0.0.1:8080/ ]
7. Repetir os passos 4, 5 (apenas o comando [iii]) e 6 para **abrir a ferramenta após a instalação**.

## 3.2 Esquema sintético de anotação

1. Ativação da função de anotação de sinalizadores na ferramenta rstWeb
2. Carregamento do arquivo "anotacao-projeto" na pasta "*signals*" da ferramenta
3. Carregamento do texto na ferramenta
4. Anotação dos sinalizadores intrassentenciais na relação
5. Associação das classificações de "Tipo" e "Subtipo" ao sinalizador
6. Indicação do(s) sinalizador(es) no texto
7. Indicação de múltiplos sinalizadores ou casos de dúvidas
8. Correção ou salvamento da anotação
9. Finalização da anotação
10. Registro da anotação

[2] Disponível em: https://youtu.be/YwEcMtbkh3U

## 3.3 Esquema detalhado de anotação

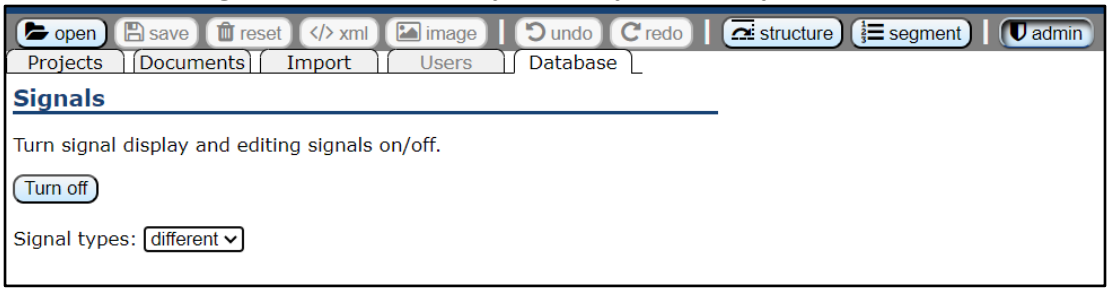
A ferramenta rstWeb (Zeldes, 2016) é uma plataforma desenvolvida para facilitar a análise e a anotação de textos com base na RST. Essa ferramenta permite aos usuários realizar análises estruturais detalhadas dos textos, identificando EDUs e suas relações de coerência conforme proposto pela teoria.

Inicialmente, a instalação da ferramenta inclui um conjunto padrão de relações RST e SDs advindos da proposta para língua inglesa (Taboada; Das, 2018; Gessler; Liu; Zeldes, 2019). No entanto, a rstWeb oferece a flexibilidade para que o usuário forneça seu próprio conjunto de relações ou SDs, permitindo uma personalização conforme as necessidades específicas de análise de textos.

### *3.3.1 Ativação da função de anotação de sinalizadores*

Na ferramenta rstWeb, clique na aba "admin" > Guia "Database" > e selecione na seção a opção "Turn on" na seção Signals. Nessa mesma seção, selecione a opção "default" dos sinalizadores (Figura 17).

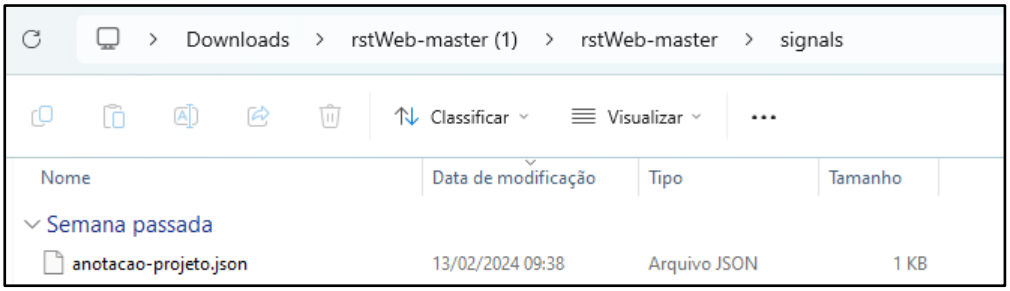
**Figura 17** - Tela de ativação da função de anotação de sinalizadores




### *3.3.2 Carregamento do arquivo "anotacao-projeto" na pasta "signals" da ferramenta*

Baixe o arquivo disponível aqui. Substitua os arquivos disponíveis na pasta '*signals*' pelo arquivo baixado (Figura 18).

**Figura 18** - Como carregar o arquivo com etiquetas na ferramenta



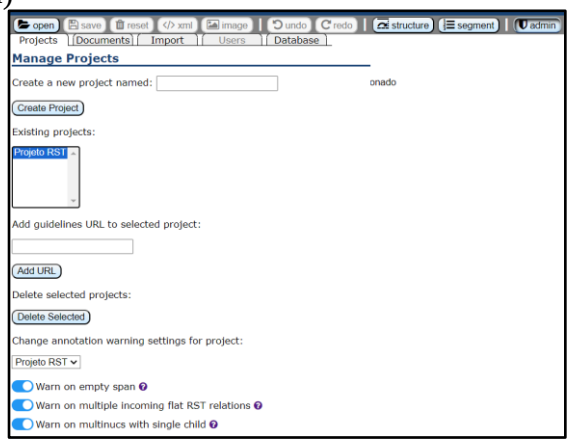
### *3.3.3 Carregamento do texto na ferramenta*

Para carregar o documento, clique em “admin” > “Create project” > Crie “Projeto RST” > Acione as configurações “Warn on empty span”, “Warn on mutiple incoming flat RST relations” e “Warn on multinucs with single child”. Importante destacar que a criação do projeto será necessária apenas na primeira vez que você realizar a anotação (Figura 19a).

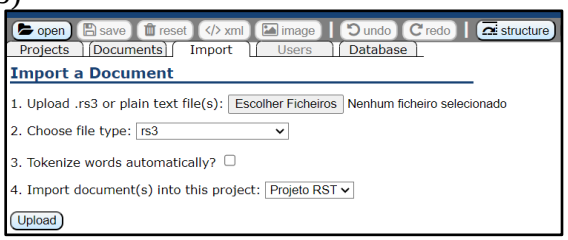
Após isso, “Import” > Escolha o documento em formato rs3 em “Escolher ficheiros” > Escolha o formato do documento para extensão .rs3 > “Upload” (Figura 19b).

**Figura 19** - Carregamento do texto na ferramenta - parte 1

a)



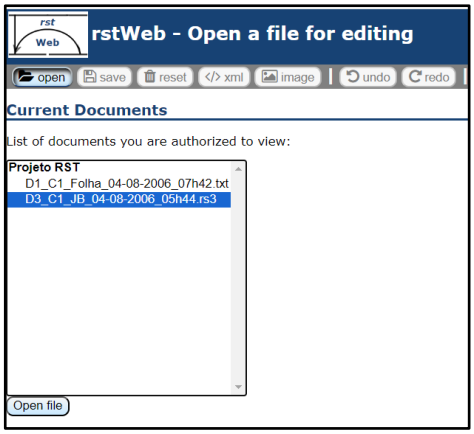

b)



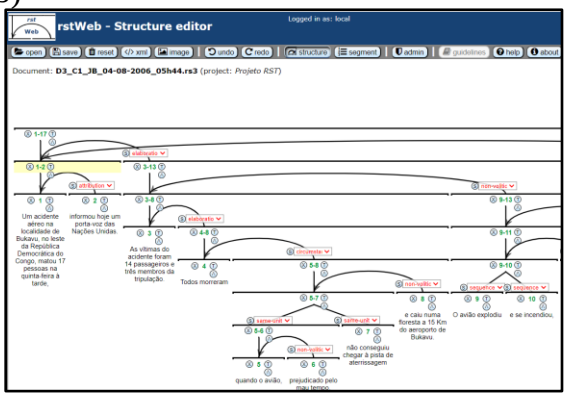

Em seguida, clique na aba “Open” > Escolha o documento. Após isso, o texto anotado será carregado na tela (Figura 20).

**Figura 20** - Carregamento do texto na ferramenta - parte 2

a)



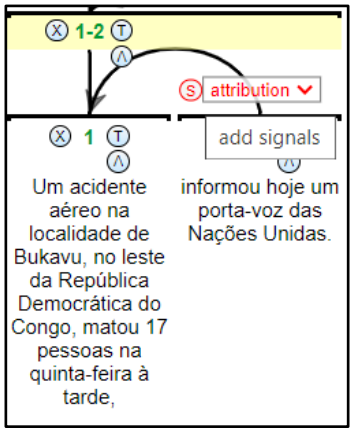

b)



### *3.3.4 Anotação dos sinalizadores na relação*

Clicar em “S”, ao lado do nome da relação RST, e será aberta a bandeja de anotação de sinal (Figura 21).

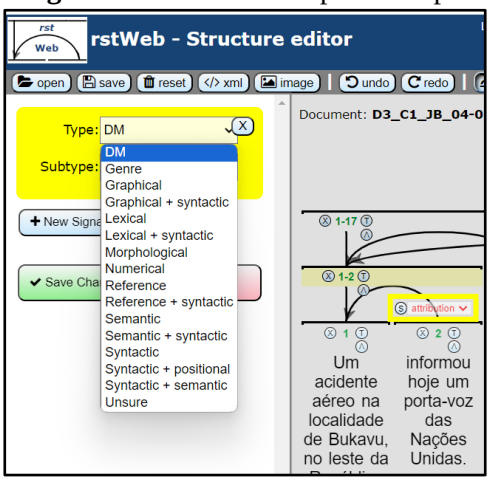
**Figura 21** - Tela de ativação da função de anotação de sinalizadores




### *3.3.5 Associação das classificações de “Tipo” e “Subtipo” ao sinalizador*

Escolher o tipo (“Type”) e o subtipo (“Subtype”) do sinalizador a partir da lista que será aberta (Figura 22).

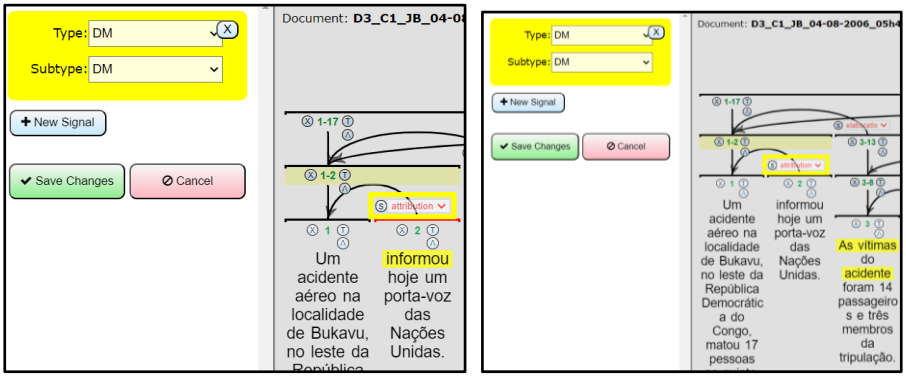
**Figura 22** - Escolha do tipo e subtipo do SD



### *3.3.6 Indicação do(s) sinalizador(es) no texto*

Indicar no texto qual o sinalizador, que pode ser uma ou mais palavras, pontuação, sinal gráfico etc, conforme a taxonomia utilizada (Figura 23).

**Figura 23** - Marcação de tokens que representam o SD de uma relação




**NOTA:** *O sinalizador a ser indicado não precisa estar contido dentro do mesmo trecho do texto da relação que você está anotando, mesmo em ocorrências intrassentenciais.*

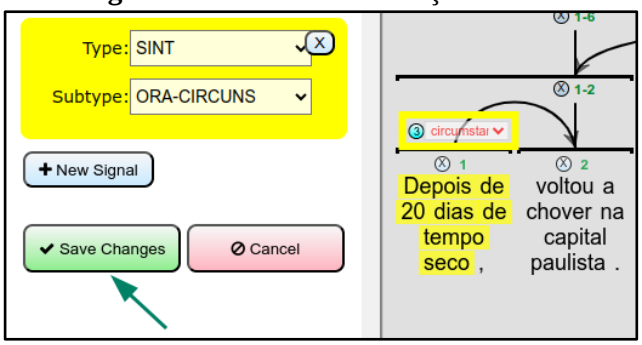
### *3.3.7 Indicação de múltiplos sinalizadores ou casos de dúvidas*

Caso haja mais de um sinalizador para aquela relação, clique em "New signal" e repita os passos 5 e 6. Caso você encontre um sinalizador que não esteja presente na lista, clique em "CPD" (Casos para discussão) e registre o novo sinalizador (*cf. Etapa 1.3.9*).

### *3.3.8 Correção ou salvamento da anotação*

Caso você queira corrigir a anotação, basta mudar o tipo e o subtipo do sinalizador, ou mesmo clicar em "X" e excluir a anotação realizada e fazer uma nova. Para salvar, clique em "Save changes" (Figura 24).

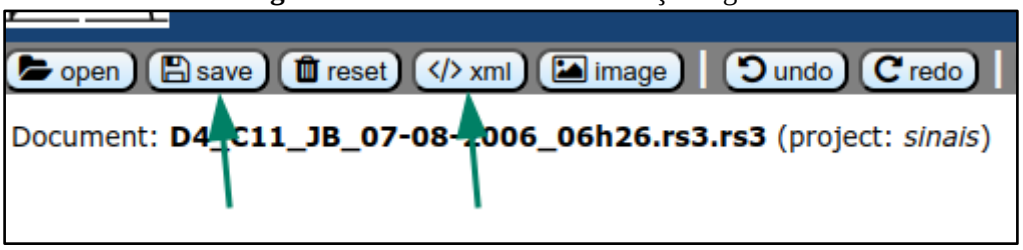
**Figura 24** - Salvando a anotação de SDs



### *3.3.9 Finalização da anotação*

Repetir as subetapas 4, 5, 6 e/ou 7 e/ou 8 até finalizar todas as relações do texto. Após isso, clique na aba “Save” e baixe o arquivo em “xml”. Na Figura 25, destaca-se os botões utilizados para essas funções.

**Figura 25** - Botões de salvar anotação e gerar XML




### *3.3.10 Registro da anotação*

Após concluir a anotação, registrar a tarefa na planilha de trabalho com todas as informações solicitadas. Além disso, enviar o arquivo xml anotado no formulário.

## Agradecimentos

Este trabalho foi realizado no *Center for Artificial Intelligence* da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP nº 2019/07665- 4) e pela IBM Corporation. O projeto também contou com apoio do Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei nº 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado como Residência no TIC 13, DOU 01245.010222 /2022-44.

Além disso, contamos com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) por meio de bolsas do Programa Institucional de Bolsas de Iniciação Científica (PIBIC) 2023 – 2024 cedidas em editais próprios da Universidade Federal da Bahia (UFBA) e da Universidade Federal de Sergipe (UFS).

## Referências

CARLSON, Lynn; MARCU, Daniel. Discourse tagging reference manual. **ISI Technical Report ISI-TR-545**, v. 54, n. 2001, p. 56, 2001.

CARDOSO, Paula CF et al. CSTnews-a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In: **Proceedings of the 3rd RST Brazilian Meeting**. 2011. p. 88-105.

DA CUNHA, Iria et al. A symbolic approach for automatic detection of nuclearity and rhetorical relations among intra-sentence discourse segments in Spanish. In: **Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part I 13**. Springer Berlin Heidelberg, 2012. p. 462-474.

DAS, Debopam; TABOADA, Maite. RST Signalling Corpus: A corpus of signals of coherence relations. **Language Resources and Evaluation**, v. 52, p. 149-184, 2018.

FRASER, Bruce. An Account of Discourse Markers. **International Review of Pragmatics**, v. 1, n. 2, p. 293–320, 2009. DOI: https://doi.org/10.1007/s10579-017-9383-x.

GESSLER, Luke; LIU, Yang Janet; ZELDES, Amir. A discourse signal annotation system for RST trees. In: **Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019**. 2019. p. 56-61.

HERNAULT, Hugo; PRENDINGER, Helmut; VERLE, David A. du; *et al*. HILDA: A Discourse Parser Using Support Vector Machine Classification. **Dialogue & Discourse**, v. 1, n. 3, p. 1–33, 2010.

LIU, Yang; ZELDES, Amir. Discourse relations and signaling information: Anchoring discourse signals in RST-DT. **Society for Computation in Linguistics**, v. 2, n. 1, 2019.

MANN, William C.; THOMPSON, Sandra A. Rhetorical Structure Theory: Toward a functional theory of text organization. **Text - Interdisciplinary Journal for the Study of Discourse**, v. 8, n. 3, 1988.

MARCU, Daniel. **The Theory and Practice of Discourse Parsing and Summarization**. 1. ed. London/England: MIT Press, 2000.

PARDO, Thiago Alexandre Salgueiro. **Métodos para análise discursiva automática**. 2005. Tese de Doutorado. Universidade de São Paulo.

PARDO, Thiago Alexandre Salgueiro ; NUNES, Maria das Graças Volpe . On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. **Revista de Informática Teórica e Aplicada**, v. 15, n. 2, p. 43–64, 2008.

SOUZA, Jackson Wilke da Cruz; CARDOSO, Paula Christina Figueira; RODRIGUES, Roana. Systematic Review of Studies on Rhetorical Structure Theory (RST)/Revisão

sistemática de estudos sobre Rhetorical Structure Theory (RST). **REVISTA DE ESTUDOS DA LINGUAGEM**, v. 31, n. 3, p. 1643-1675.

RODRIGUES, Roana; SOUZA, Jackson Wilke; CARDOSO, Paula Christina Figueira. Sinalizadores retórico-discursivos: revisitando a anotação RST no córpus CSTNews. In: **Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. SBC, 2023. p. 249-257.

TABOADA, Maite; DAS, Debopam. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. **Dialogue & Discourse**, v. 4, n. 2, p. 249-281, 2013.

TABOADA, Maite; MANN, William C. Rhetorical Structure Theory: looking back and moving ahead. **Discourse Studies**, v. 8, n. 3, p. 423–459, 2006.

ZELDES, Amir. rstWeb - A Browser-based Annotation Interface for Rhetorical Structure Theory and Discourse Relations. *In*: **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016.