

SCC0633/SCC5908

PROCESSAMENTO DE LINGUAGEM NATURAL



EX_MACHINA: INSTINTO ARTIFICIAL (2004)



O TESTE DE TURING E A REAL CAPACIDADE LINGUÍSTICA

A HISTÓRIA, AS EXPECTATIVAS E A REALIDADE

De Eliza

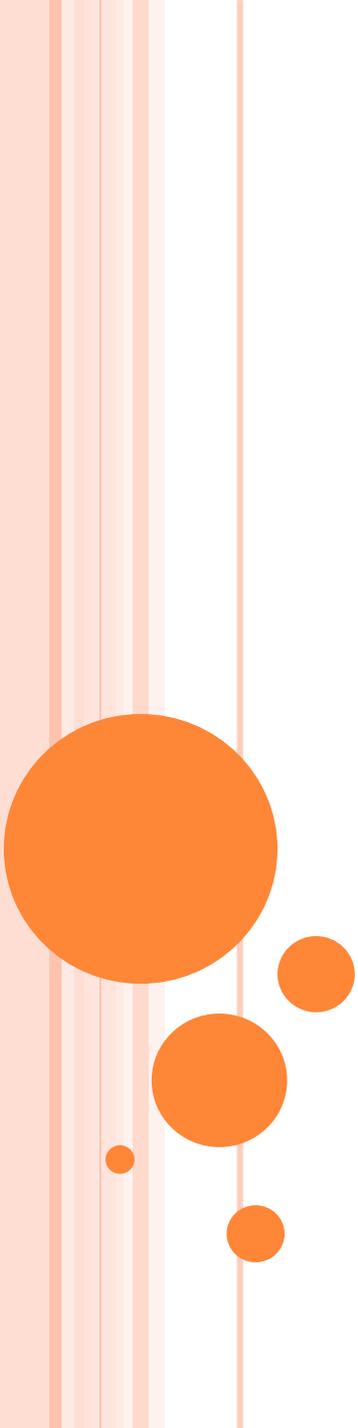
aos embeddings

Passamos pelo teste de Turing?

(e o que a leitura da semana ensinou)

Temos a solução de todos os
problemas?

(e o que a história ensinou)



ANOTAÇÃO DE CÓRPUS

SCC5908 Introdução ao Processamento de Língua Natural

SCC0633 Processamento de Linguagem Natural

Prof. Thiago A. S. Pardo

RELEMBRANDO

- Métodos de contagem e Zipf
- Estatística, Bayes e modelo *noisy-channel*
- Hipótese distribucional, semântica e vetores
- De representações *bag-of-words* aos *embeddings* modernos
- Word2vec, BERT

APRENDENDO A PARTIR DE CÓRPUS

- Modelos matemáticos, estatísticos e distribucionais são muito úteis atualmente
 - Caracterização de fenômenos
 - Padrões e probabilidades
 - Níveis de significado (implícito)
 - Etc.
- Mas muitas tarefas de PLN ainda precisam de outras anotações linguístico-computacionais sofisticadas, possivelmente de outros níveis
 - Sintaxe
 - Semântica (explícita)
 - Discurso
 - Etc.

VAMOS PENSAR

- Como faríamos para desenvolver um sistema que detecta emoções em tweets?
 - Até podemos usar os *embeddings* das palavras para ajudar no “treinamento” do sistema, mas há muitas questões a resolver antes
 1. Que dados usaremos? Podemos usá-los? Qual a distribuição dos fenômenos de interesse?
 2. Que modelo de emoções usaremos? Há modelos de emoções?
 3. Com que anotadores contaremos? Como treiná-los e garantir um bom desempenho na anotação?
 4. Como será o procedimento de anotação?
 5. Qual interface de anotação será utilizada?
 6. Como avaliar a qualidade e confiabilidade da anotação realizada?
 7. Onde disponibilizar os dados anotados?

ANOTAÇÃO DE CÓRPUS

- Também é ciência (Hovy e Lavid, 2010)
 - Questões de pesquisa teóricas e práticas
 - Questões éticas
 - Questões legais
- **Córpus na história**
 - Grandes investimentos no passado
 - Maior necessidade de aplicação prática no presente
 - A relação com Aprendizado de Máquina
 - A explosão da Ciência de Dados e a valorização dos dados
 - Córpus vs. dataset

Questões de pesquisa

7 questões principais (Hovy e Lavid, 2010)

1. Seleção do corpus (e as questões já estudadas anteriormente)

- Copyright
- Disponibilidade
- Gênero textual
- Domínio
- Tempo
- Representatividade
- Balanceamento
- Web como corpus
- Etc.

Como garantir representatividade e balanceamento?

WebCorp: The Web as Corpus

www.webcorp.org.uk/live/

The screenshot shows the WebCorp Live website interface. At the top, the logo "WebCorp Live" is displayed in a teal-to-white gradient, with the tagline "Concordance the web in real-time." below it. A dark teal navigation bar contains the following links: "Search" (highlighted in orange), "Wordlist Tool", "User Guide", "WebCorp LSE", "Publications", and "Feedback".

The main content area features a descriptive paragraph: "WebCorp Live lets you access the Web as a corpus - a large collection of texts from which examples of real language use can be extracted. [More...](#)"

Below this is a search form with the following fields and options:

- Search:** A text input field with an information icon (i) to its right.
- Case Insensitive:** A checkbox that is checked.
- Span:** A dropdown menu set to "50 characters" with an information icon (i) to its right.
- Search Engine:** A dropdown menu set to "FAROO".
- Language:** A dropdown menu set to "Not specified" with an information icon (i) to its right.

Below the search form is a section titled "Advanced Options" in teal text. At the bottom of the form are two orange buttons: "Redefinir" and "Search".

On the right side of the page, there is a grey-bordered box with the following content:

- WebCorp** Linguist's Search Engine
- Have you tried WebCorp LSE?
- Our large-scale search engine with more search options, part-of-speech tags and quantitative analyses.
- [More details...](#)

Questões de pesquisa

7 questões principais (Hovy e Lavid, 2010)

2. Instanciação da teoria

- Categorias/etiquetas
- Esquema de anotação
- Diretrizes
- Manual
- Concordância
- “Neutralização” da teoria
- Etc.

Qual o limite da neutralização?
Pode “robotizar” a tarefa e matar
a subjetividade humana?

Questões de pesquisa

7 questões principais (Hovy e Lavid, 2010)

3. Seleção e treinamento de anotadores

- Expertise dos anotadores
- Quantidade de anotadores
- Pagamento, recompensa
- Risco de supertreino
- Pouco treino
- Falta de consistência
- Etc.

Pode só 1 anotador? Em que situações?

Questões de pesquisa

7 questões principais (Hovy e Lavid, 2010)

4. Especificação do procedimento de anotação

- Anotações preliminares
- Discussão de discordâncias
- Presença de juiz
- Quantidade de “passadas de anotação”
- Planejamento, cronograma de execução
- Questões humanas (cansaço, vontade, erro, inconsistência)
- Etc.

Questões de pesquisa

7 questões principais (Hovy e Lavid, 2010)

5. Projeto da interface de anotação

- Facilidade
- Sem *bias*
- Ordem de opções de anotação
- Disposição na tela
- Fatores humanos
- Etc.

Questões de pesquisa

7 questões principais (Hovy e Lavid, 2010)

6. Escolha e aplicação de medidas de avaliação

- Concordância esperada e o propósito da anotação
- Medidas simples e mais sofisticadas de concordância
- Estabilidade e consistência do anotador
- Etc.

Questões de pesquisa

7 questões principais (Hovy e Lavid, 2010)

7. Disponibilização e manutenção do produto

- Reuso
- Licença
- Portais de córpus
- Etc.

Uso interno vs. disponibilização pública

Córpus

Criação e anotação

Tendências

Ferramentas pré-fabricadas

Ferramentas customizáveis

- *MetaAnn: Um Gerador de Ferramentas para Anotação de Textos* (Missão e Roman, 2013)

Questões

- Custo de customização?
- A mais apropriada para a tarefa?

The screenshot displays the MetaAnn software interface. At the top, a window titled "Documento" contains a text block: "Tinha-me lembrado a definição que José Dias dera deles, 'olhos de cigana oblíqua e dissimulada.' Eu não sabia o que era oblíqua, mas dissimulada sabia, e queria ver se podiam chamar assim. Capitu deixou-se fitar e examinar. Só me perguntava o que era, se nunca os vira, eu nada achei extraordinário; a cor e a doçura eram minhas conhecidas."

Below the document, there is a section for "ID da Unidade" with the value "0001" and "Unidade" with the text "Tinha-me lembrado a definição que José Dias dera deles".

The "Classificação" section features five categories, each with a dropdown menu labeled "ESCOLHA..." and a corresponding text input field with a "Limpar" button:

- CATEGORIA 1: ESCOLHA... | OPÇÃO X | Limpar
- CATEGORIA 2: ESCOLHA... | OPÇÃO C, OPÇÃO A, OPÇÃO B, OPÇÃO D, | Limpar
- CATEGORIA 3: ESCOLHA... | OPÇÃO 2, OPÇÃO 1, | Limpar
- CATEGORIA 4: | Text Here ...
- CATEGORIA 5: |

At the bottom, a navigation bar includes buttons for navigation (left and right arrows) and actions: "Salvar", "Ir", and "Sair".

Ferramentas pré-fabricadas

Ferramentas customizáveis

- *AgreeCalc: Uma Ferramenta para Análise da Concordância entre Múltiplos Anotadores* (Alvares e Roman, 2013)

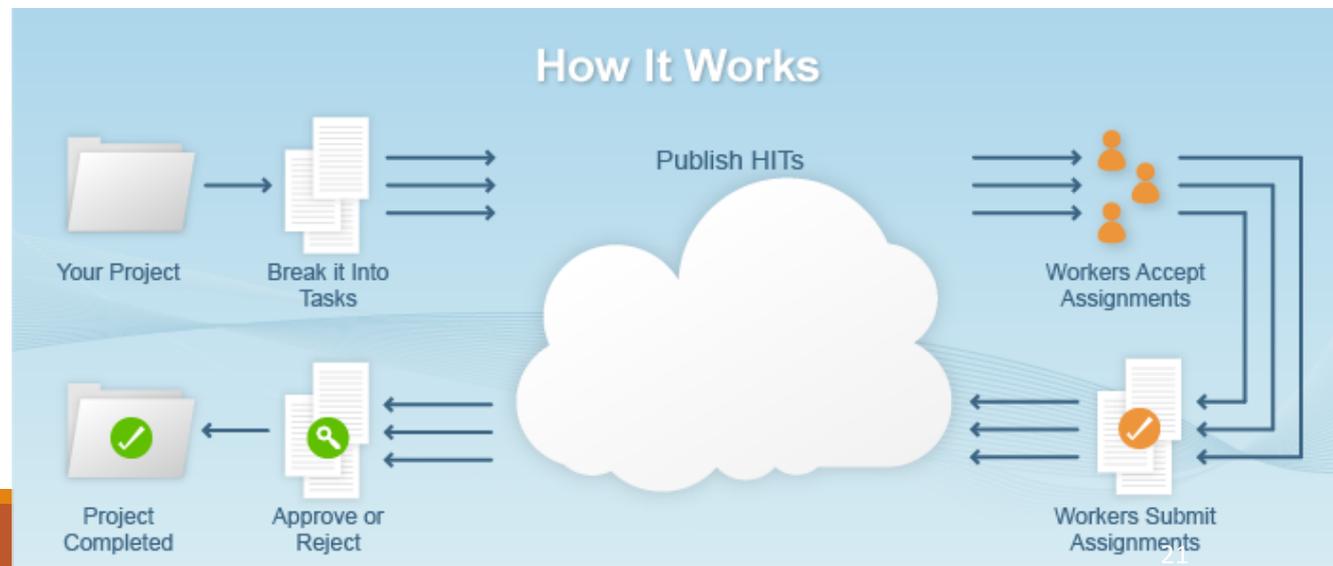


Crowdsourcing

Amazon Mechanical Turk

Questões (Fort et al., 2011)

- Custo
- Qualidade vs. quantidade de anotadores, confiabilidade
- Ética



Colaboração

Anotação colaborativa, distribuída (escopo normalmente menor do que *crowdsourcing*, com mais “controle”)

Questões

- Seleção de anotadores, treinamento
- Controle de execução da tarefa: qualidade, tempo

Vantagens

- Dados
- Ferramentas



Description

Collective Elaboration of a Coreference Annotated Corpus for Portuguese Texts

Organizers

- Evandro Fonseca, PUCRS Porto Alegre
- Vinicius Sesti, PUCRS Porto Alegre
- Ana Luisa Leal, UMAC Macau
- Sandra Collovini, PUCRS Porto Alegre
- Renata Vieira, PUCRS Porto Alegre
- Paulo Quaresma, UEVORA Évora

Anotação semiautomática

Uso de ferramentas automáticas de anotação

+

Revisão humana

Revisar normalmente é mais fácil do que anotar do zero

- Alta influência do tipo de anotação

Córpus

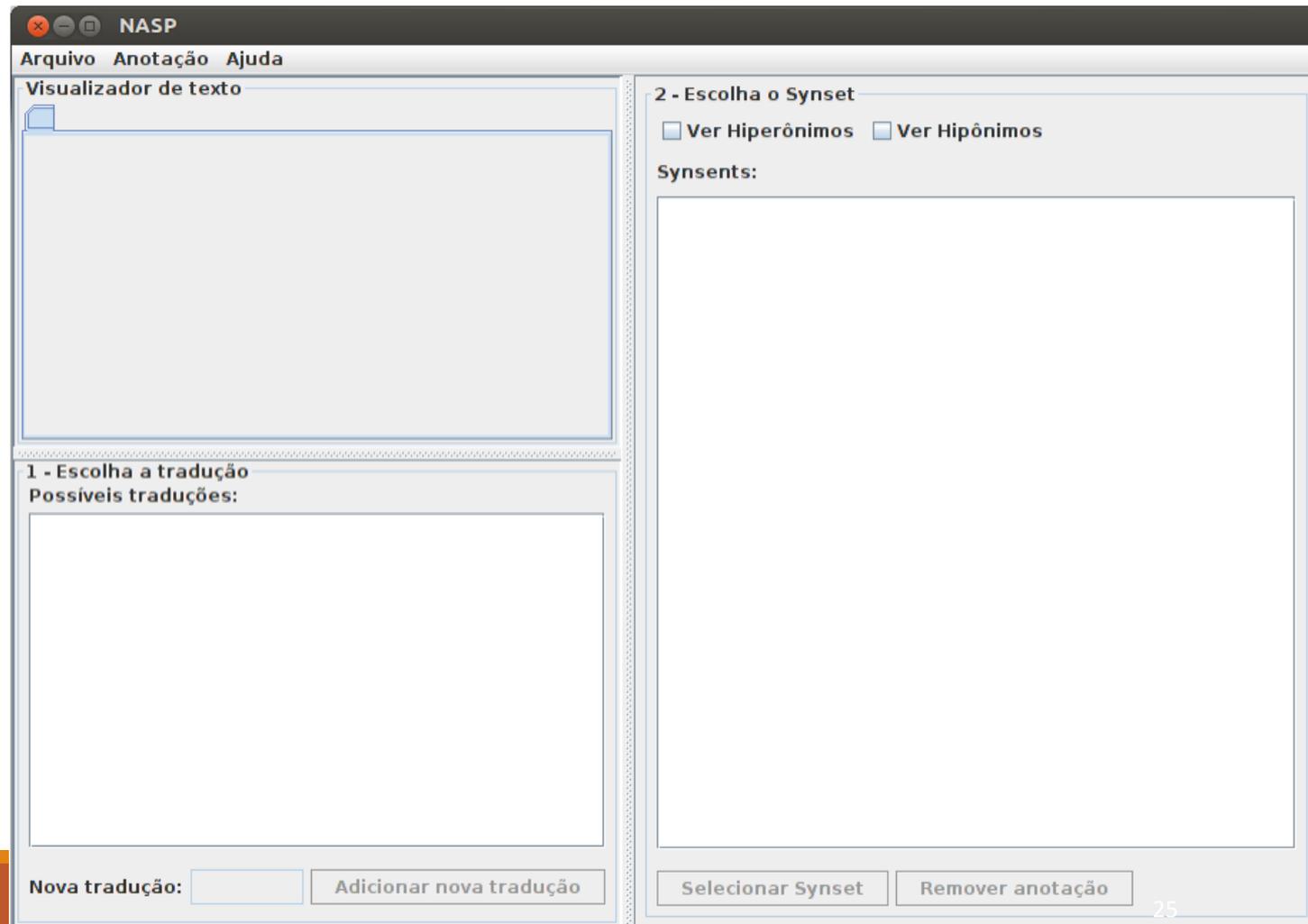
Criação e anotação

Exemplos reais

Sentido de palavras

NASP

Nóbrega
(2013)



Sentido de palavras

NASP

Nóbrega
(2013)

The screenshot displays the NASP (Network Analysis and Semantic Processing) software interface. The window title is "NASP" and it has a menu bar with "Arquivo", "Anotação", and "Ajuda".

The main area is divided into two panes:

- Visualizador de texto (Text Viewer):** Shows a text document with three tabs: "Texto 1", "Texto 2", and "Texto 3". The text in "Texto 1" is: "ram após a queda de um avião de passageiros na ONU, o avião, de fabricação russa, estava tentando pousar numa montanha e caiu, em chamas, sobre uma floresta densa no Congo, onde 51 companhias privadas operam voos regulares. O avião, do tipo Boeing 737, foi operado pela Air Traset, levava 14 passageiros e três tripulantes. O acidente ocorreu numa montanha de Lugushwa em direção a Bukavu, numa distância de 100 metros do aeroporto. Atualmente para transporte na República Democrática do Congo". The words "avião" and "passageiros" are highlighted with red boxes.
- 2 - Escolha o Synset (Choose the Synset):** A list of synsets for the word "avião". Each synset includes a set of related terms, a definition, and an example. The synsets are:
 - 1 -- {airplane, aeroplane, plane} an aircraft that has a fixed wing and is powered by propulsion system. Ex. 1: "the flight was delayed due to trouble with the engine."
 - 2 -- {plane, sheet} (mathematics) an unbounded two-dimensional shape that has no thickness, zero curvature, and is perfectly flat. Ex. 1: "we will refer to the plane of the graph as the xy-plane." Ex. 2: "any line joining two points on a plane lies in the plane."
 - 3 -- {plane} a level of existence or development. Ex. 1: "he lived on a worldly plane."
 - 4 -- {plane, planer, planing machine} a power tool for smoothing or shaping wood.
 - 5 -- {plane, carpenter's plane, woodworking plane} a carpenter's hand tool with an adjustable blade for smoothing or shaping wood. Ex. 1: "the cabinetmaker used a plane for the first time."

At the bottom of the interface, there are two main sections:

- 1 - Escolha a tradução (Choose the translation):** Shows "Possíveis traduções:" (Possible translations) with a list containing "plane" and "airplane". Below this is a "Nova tradução:" (New translation) field and an "Adicionar nova tradução" (Add new translation) button.
- 2 - Escolha o Synset:** At the bottom of this pane are two buttons: "Selecionar Synset" (Select synset) and "Remover anotação" (Remove annotation).

Sentido de palavras

NASP

Nóbrega
(2013)

The screenshot shows the NASP software interface. The main window is titled "NASP" and has a menu bar with "Arquivo", "Anotação", and "Ajuda". The central area is a "Visualizador de texto" (Text Viewer) with three tabs: "Texto 1", "Texto 2", and "Texto 3". The text in "Texto 1" is: "ram após a queda de um avião de passageiros na F... ONU, o avião, de fabricação russa, estava tentand... uma montanha e caiu, em chamas, sobre uma flores... entes no Congo, onde 51 companhias privadas oper... do pela Air Traset, levava 14 passageiros e três trip... ineira de Lugushwa em direção a Bukavu, numa dist... amente para transporte na República Democrática d...". The word "avião" is highlighted in yellow, and "passageiros" is highlighted in red. Below the text viewer is a section titled "1 - Escolha a tradução" (Choose the translation) with the sub-heading "Possíveis traduções:" (Possible translations:). It lists "plane" and "airplane". At the bottom of this section are two buttons: "Nova tradução:" (New translation:) followed by an empty text box, and "Adicionar nova tradução" (Add new translation). To the right of the text viewer is a section titled "2 - Escolha o Synset" (Choose the Synset). It has two checkboxes: "Ver Hiperônimos" (View Hyperonyms) and "Ver Hipônimos" (View Hyponyms). Below these is a list of "Synsets":

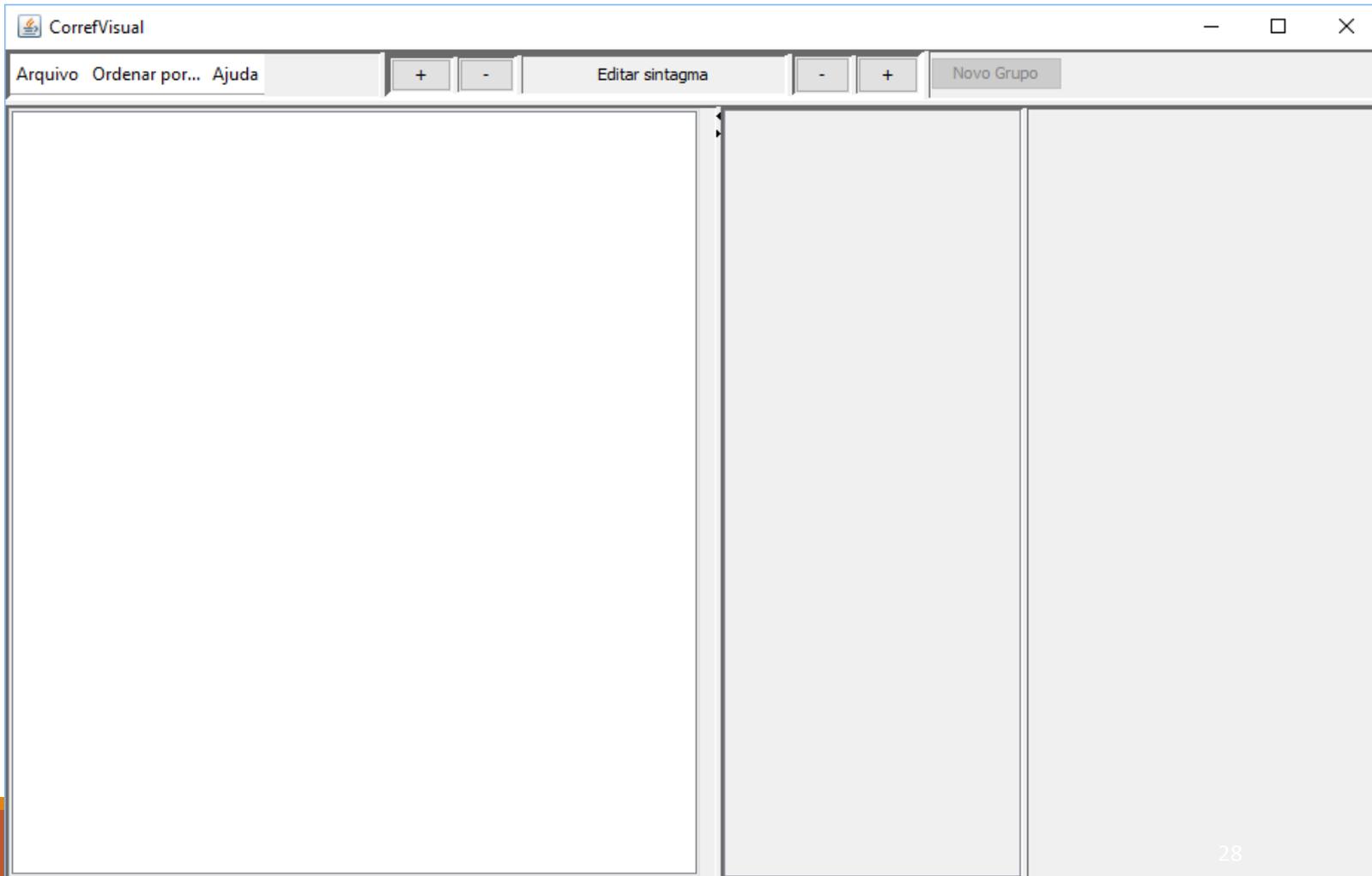
- 1 -- {airplane, aeroplane, plane} an aircraft that has a fixed wing and is powered by pi...
 - Ex. 1: "the flight was delayed due to trouble w..."
- 2 -- {plane, sheet} (mathematics) an unbounded two-dimensional shape...
 - Ex. 1: "we will refer to the plane of the graph a..."
 - Ex. 2: "any line joining two points on a plane li..."
- 3 -- {plane} a level of existence or development
 - Ex. 1: "he lived on a worldly plane"
- 4 -- {plane, planer, planing machine} a power tool for smoothing or shaping wood
- 5 -- {plane, carpenter's plane, woodworking plane} a carpenter's hand tool with an adjustable blade for...
 - Ex. 1: "the cabinetmaker used a plane for the f..."

At the bottom of this section are two buttons: "Selecionar Synset" (Select Synset) and "Remover anotação" (Remove annotation). The page number "27" is visible in the bottom right corner.

Cadeias de correferência

CorrefVisual

Fonseca et
al. (2017)



Cadeias de correferência

CorrefVisual

Fonseca et al. (2017)

The screenshot shows the CorrefVisual application window. The main text area contains a paragraph about the common origin of various dog breeds. Several phrases are highlighted in blue, and a sidebar on the right lists related terms under different categories.

Main Text:

Seja **um poodle ou um pitbull , um cocker spaniel ou um dinamarquês , o seu cachorro e os de os seus amigos** têm uma origem comum : **todos eles** descendem de lobos domesticados por a primeira vez em o leste de a Ásia , há possivelmente meros 15 mil anos . Meros porque esse período é muito curto em termos de evolução . Os primeiros animais aparentados com **os cachorros modernos** os primeiros canídeos , foram identificados em fósseis de 37 milhões de anos achados em a América_do_Norte e aparecem em a forma de fósseis em a Europa há 7 milhões de anos . Apesar_de os lobos já estarem em o continente americano há dezenas de milhares de anos , **os cachorros modernos** só fizeram a viagem de volta de a Eurásia a as Américas entre 12 mil e 14 mil anos atrás , acompanhando levadas migratórias humanas . Dois estudos sobre o material genético de os cães publicados hoje em a revista Science (www.sciencemag.org) revelaram que , diferentemente de o que indicavam os achados arqueológicos , os primeiros cães domésticos surgiram em o Extremo_Oriente , e não em o Oriente_Médio . E que , apesar_da a sua longa permanência em as Américas , a grande maioria de os cães de o continente veio mesmo de a Europa depois de a descoberta de Cristóvão_Colombo , em 1492 . Um de os estudos , feito por cientistas suecos e chineses , envolveu a análise de o DNA de cachorros de várias partes de o planeta . A equipe de Peter_Savolainen , de o Instituto_Real_de_Tecnologia , de Estocolmo , Suécia , constatou que a diversidade genética era maior entre os cães de o leste asiático . Isso indicaria que a domesticação ali seria mais antiga , apesar_de mais fósseis de cães domésticos terem sido achados em o Oriente_Médio . Como também havia lobos selvagens em as Américas antes de a chegada de os humanos , outros cientistas quiseram testar as hipóteses de como e quando teria ocorrido a domesticação em o novo continente . Teria havido uma domesticação independente , ou os cães vieram junto com as migrações humanas via estreito de Bering , entre Rússia e Alasca? . Cães indígenas . Jennifer_Leonard , de a Universidade_da_Califórnia em Los_Angeles , e colegas de instituições de o Peru e de o México procuraram verificar o material genético de raças de cachorro sabidamente anteriores a a chegada de os europeus . A análise levou a duas conclusões : primeiro , as raças indígenas são tão parecidas com as asiáticas que permitem sugerir que os cães tinham chegado

Sidebar Categories and Items:

- NATUREZA**
 - um poodle ou um pitbull , um cocker spaniel
 - um cocker spaniel
 - um dinamarquês
 - o seu cachorro
 - os seus amigos
 - todos eles
 - os cachorros modernos
 - os cachorros modernos
- NATUREZA**
 - Os primeiros animais aparentados com os ca
 - os primeiros canídeos
 - os primeiros cães domésticos
 - os cães
 - os cães
 - cães domésticos
 - os cães
 - os cães
- ORGANIZAÇÃO|LOCAL**
 - o continente americano
 - as Américas
 - as Américas
 - o continente
 - as Américas
 - o novo continente
 - o continente
 - a América
- SUBSTÂNCIAS**
 - o material genético
 - o DNA
 - o material genético de raças de cachorro sa
 - o DNA
 - uma estrutura
 - as células
 - as mitocôndrias , que
 - que

Panel auxiliar (Right):

- Menções únicas
- termos
- evolução
- anos
- a América de o Norte
- anos
- que
- a sua longa permanência
- a descoberta
- Cristóvão Colombo
- o Instituto Real de Tecnologia
- Estocolmo
- Suécia
- as hipóteses de como
- Rússia
- a Universidade de a Califórnia
- instituições
- o México
- a duas conclusões
- que
- crúzamentos seletivos
- os últimos 500 anos
- particularmente úteis a os primitivos caçadores-coletores
- as mães
- a primeira vez

Análise semântica

Abstract Meaning Representation (Banarescu et al., 2013)

AMR Editor guest
Written by Ulf Hermjakob, USC/ISI Version 1.7.4z120 Nov. 11, 2018 AMR Editor URL: <https://www.isi.edu/~ulf/amr/AMR-editor.html>

(m / morrer
 :ARG1 (e / ele))

Enter text command: [QuickRef](#)

Last command: m :ARG1 ele

Or select an action template: [top](#) [add](#) [add-ne](#) [replace](#) [delete](#) [move](#) [undo](#) [load](#) [props](#) [save](#) [save as](#) [c](#)

More: [check](#) [copy](#) [dict](#) [diff](#) [generate](#) [guidelines](#) [logout](#) [meetings](#) [NE types](#) [roles](#) [search](#) [videos](#) [wil](#)

Action template [add](#) head var: role: arg:

Log: initialized empty AMR
Create new AMR_editor cookie ID 20190325.17340949.79352443
For role checking, loaded 151 roles and 13 non-roles.
For OntoNotes frame availability check, loaded 9766 verbs.

Polaridade de *tweets*

Brum (2017)

Estranho ver Nanini e Rosi Campos juntos num mesmo programa. Por muito tempo achei que eram a mesma pessoa #Encontro

Negativo

Neutro

Positivo

Não tenho certeza

A banda malta já não era lá essas coisas conseguiu ficar pior meu deus #AltasHoras

Negativo

Neutro

Positivo

Não tenho certeza

#DomingoLegal Tá Muito Bom

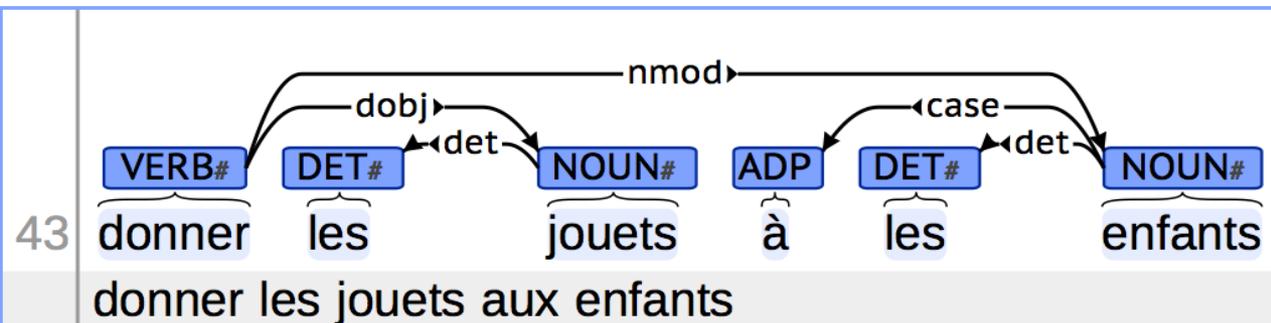
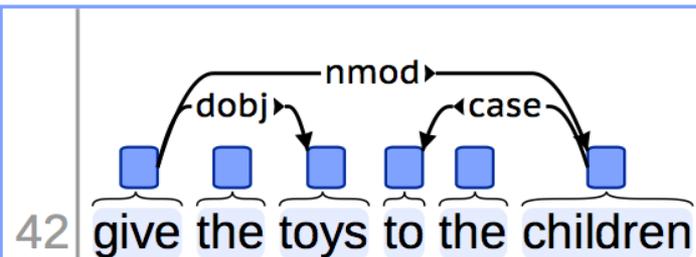
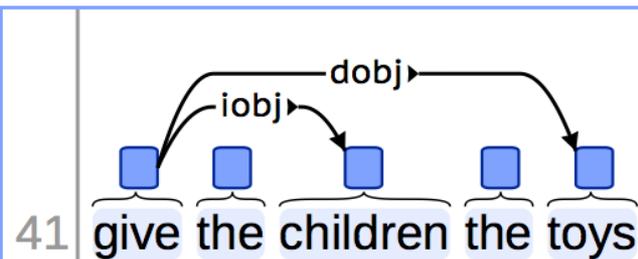
Negativo

Neutro

Positivo

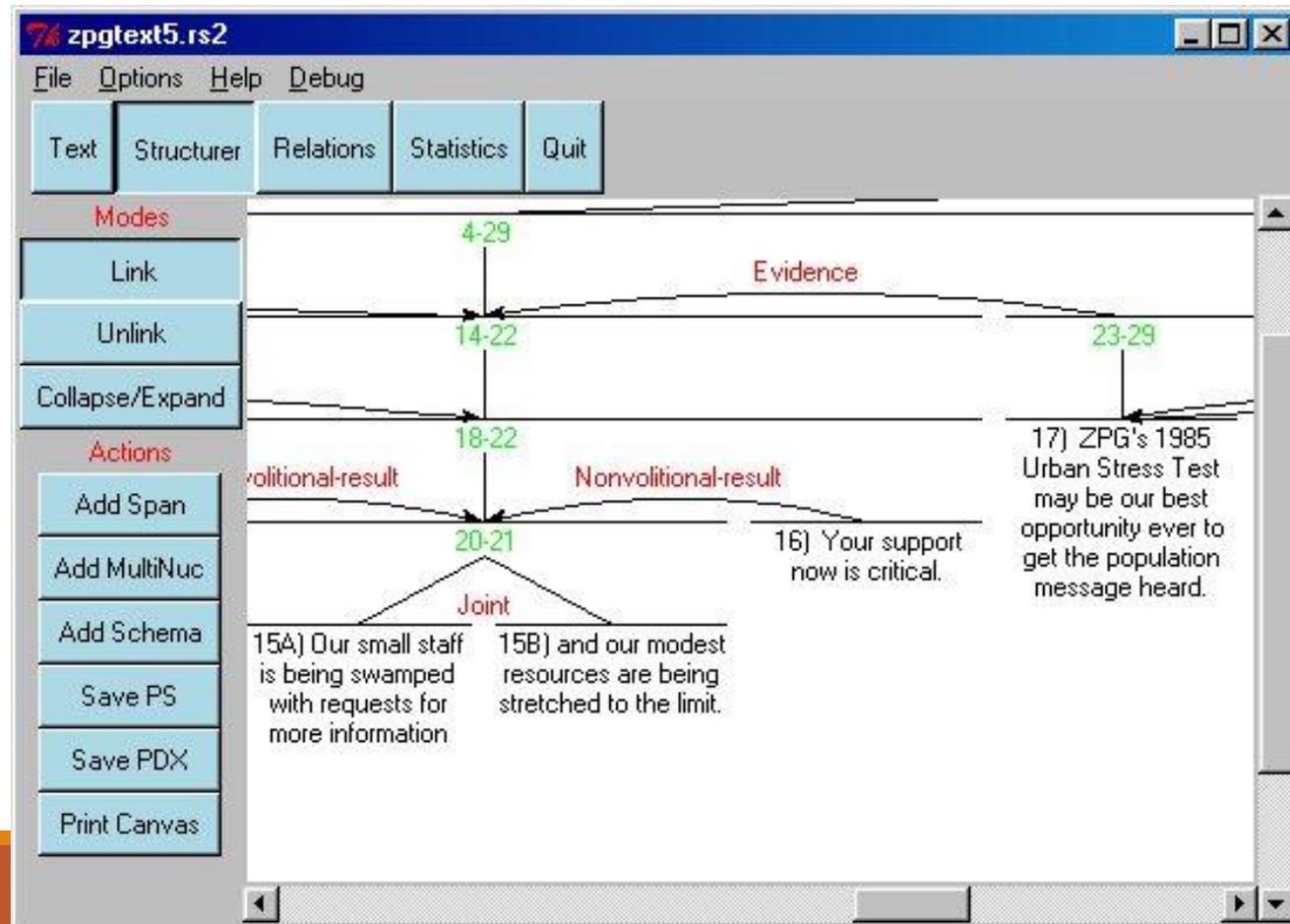
Dependências sintáticas

brat rapid annotation tool (Stenetorp et al., 2012)



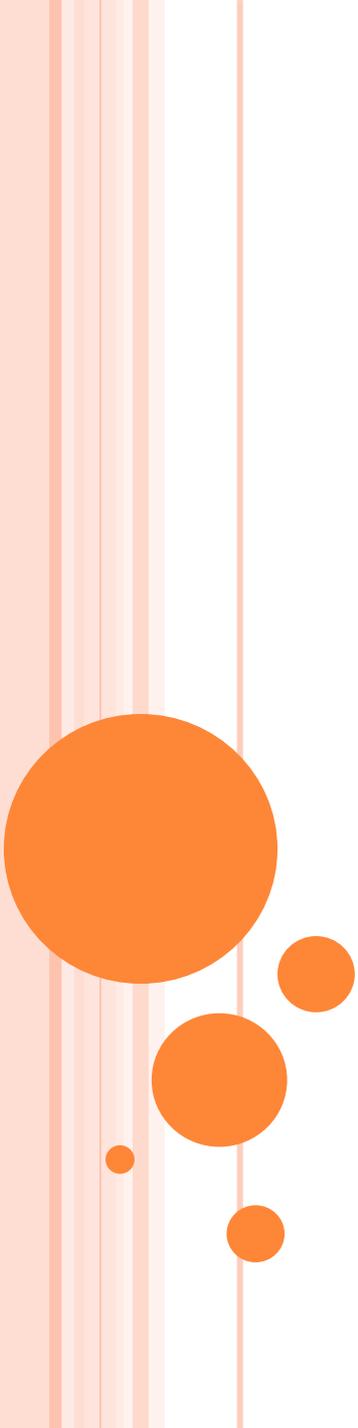
Relações retóricas

RSTTool (O'Donnell, 1997)



Caso real

O PROJETO SUCINTO



info
sucinto

summarization
for clever information access →





summarization
for clever information access →

○ www.icmc.usp.br/~tasparado/sucinto

- Projeto longo: de 2009 a 2017
 - Financiamento FAPESP, principalmente
- Foco em ***sumarização multidocumento***

○ 3 principais objetivos

- Modelagem do processo de sumarização com uso de modelos discursivos, ontologias, modelos estatísticos, etc.
- Investigação de tarefas correlatas: análise discursiva, detecção topical, resolução temporal, resolução de correferências, alinhamento, processamento multilíngue, etc.
- Caracterização linguística de sumários e de sua produção manual

DESAFIOS MULTIDOCUMENTO

○ Desafios “operacionais”

- Sumarização de 2 textos... até milhares de textos!
- Taxas de compressão muito mais altas
- Visualização textual e gráfica
- Navegação entre textos e sumários
 - História da informação e sua origem
- Organização de informações e textos
- Integração a ferramentas de busca e processamento textual

DESAFIOS MULTIDOCUMENTO

○ Desafios “linguístico-computacionais”

- Evolução de eventos no tempo
- Narração dos eventos com diversos estilos, perspectivas diferenciadas e em momentos variados
- Fontes tendenciosas, parciais, “corruptas”
 - Qualidade da informação é importante!
 - Um sumarizador deveria propagar um boato?
- Diferentes focos sobre uma mesma informação central
- Expressões referenciais diferentes, resolução de correferências multidocumento

DESAFIOS MULTIDOCUMENTO

○ Desafios “linguístico-computacionais”

- Fenômenos multidocumento
 - Informação redundante
 - Informações complementares
 - Informações contraditórias
 - Evolução de um evento, com relatos parciais ou em momentos diferentes
 - Erros
 - Discordâncias e perspectivas diferentes
- Ordenação das informações
- Coerência e coesão, por fim

DESAFIOS MULTIDOCUMENTO

○ Desafios “correlatos”

- Agrupamento de textos
 - Buscadores web
- Categorização de passagens textuais, segmentação e identificação topical
- Rotulação dos grupos de documentos
 - Termos dos textos ou elementos semânticos/ontológicos
- Etc.



summarization
for clever information access →

○ www.icmc.usp.br/~tasparado/sucinto

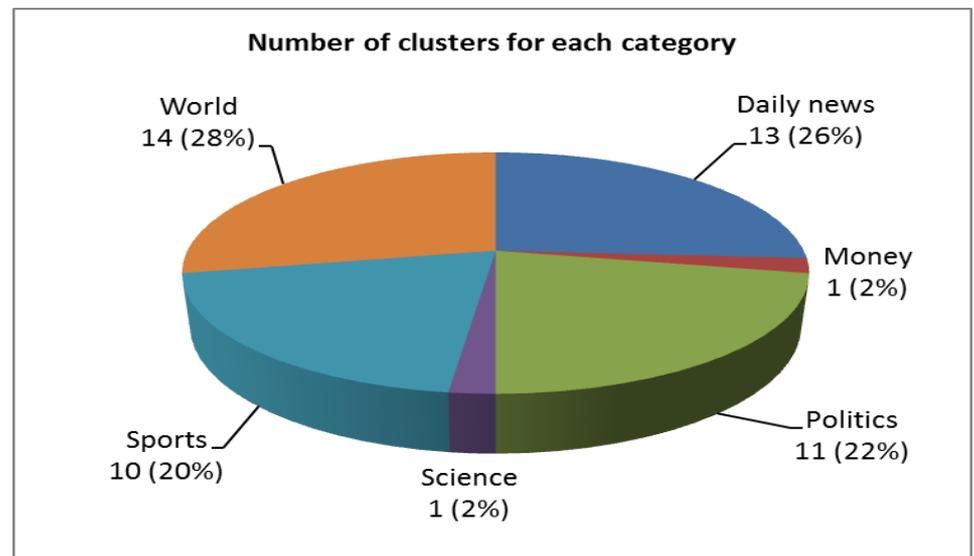
- Projeto longo: de 2009 a 2017
 - Financiamento FAPESP, principalmente
- Foco em **sumarização multidocumento**
 - Muitas vontades, mas...
 - Somente sumarização monodocumento para português
 - Sem ferramental para o português
 - **Sem córpus para o português**



O CÓRPUS

○ Criação e anotação do corpus CSTNews

- 50 grupos de textos
 - 2 ou 3 textos por grupo
 - 140 textos, no total



Aleixo, P. and Pardo, T.A.S. (2008). CSTNews: Um Corpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory). Technical Report, Universidade de São Paulo, N. 326. São Carlos-SP, May, 12p.

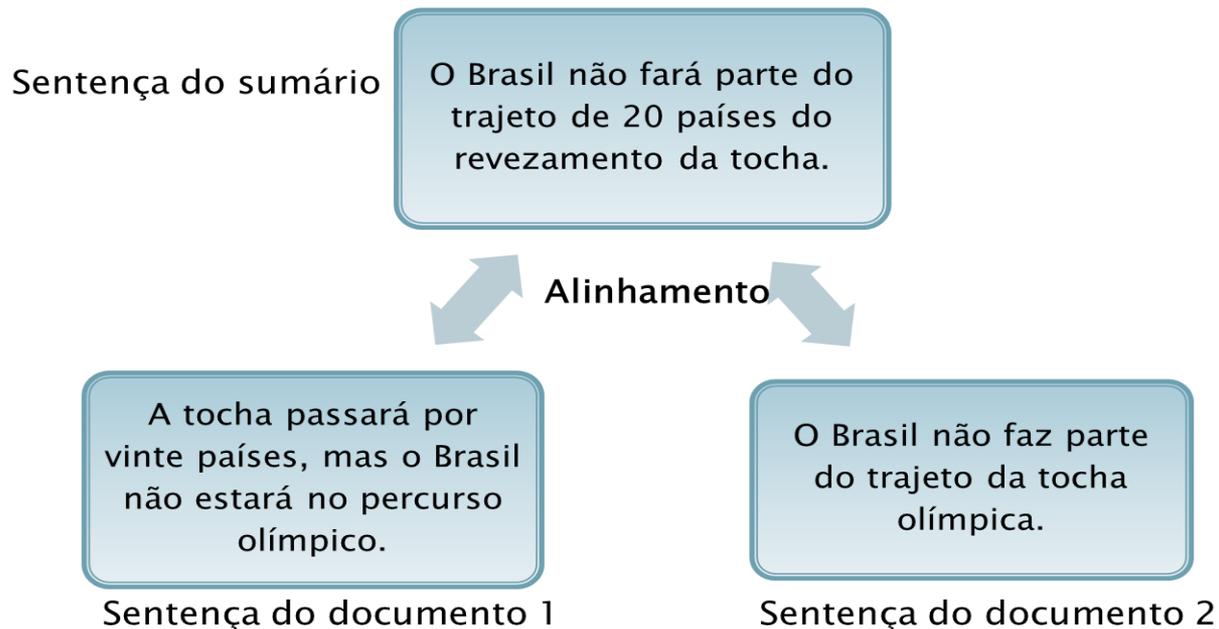
SUMÁRIOS, LÓGICO!

- Criação e anotação do córpus CSTNews
 - Sumários manuais, produzidos em etapas
 - Para cada grupo de textos
 - 1 sumário monodocumento abstrativo por texto
 - 6 sumários multidocumento abstrativos
 - 6 sumários multidocumento extrativos
 - Sumários automáticos produzidos por sistemas variados para o português

Dias, M.S.; Bokan Garay, A.Y.; Chuman, C.; Barros, C.D.; Maziero, E.G.; Nobrega, F.A.A.; Souza, J.W.C.; Sobrevilla Cabezado, M.A.; Delege, M.; Castro Jorge, M.L.R.; Silva, N.L.; Cardoso, P.C.F.; Balage Filho, P.P.; Lopez Condori, R.E.; Marcasso, V.; Di Felippo, A.; Nunes, M.G.V.; Pardo, T.A.S. (2014). Enriquecendo o Corpus CSTNews - a Criação de Novos Sumários Multidocumento. In the (on-line) Proceedings of the I Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish - ToRPorEsp, pp. 1-8. October 9. São Carlos-SP/Brazil.

ALINHAMENTO

- Origem das informações nos sumários
 - 2 anotadores, usando o Microsoft Word!
 - 2 meses de anotação
 - Foco mais linguístico



ALINHAMENTO

- Nem sempre tão claro

Sumário: Vários moradores e turistas nas regiões, inclusive brasileiros, foram retirados dos locais, enquanto outros estão se preparando para a passagem do furacão.

Documento: Na Jamaica, muitos estocaram alimentos, água, lanternas e velas.

ALINHAMENTO

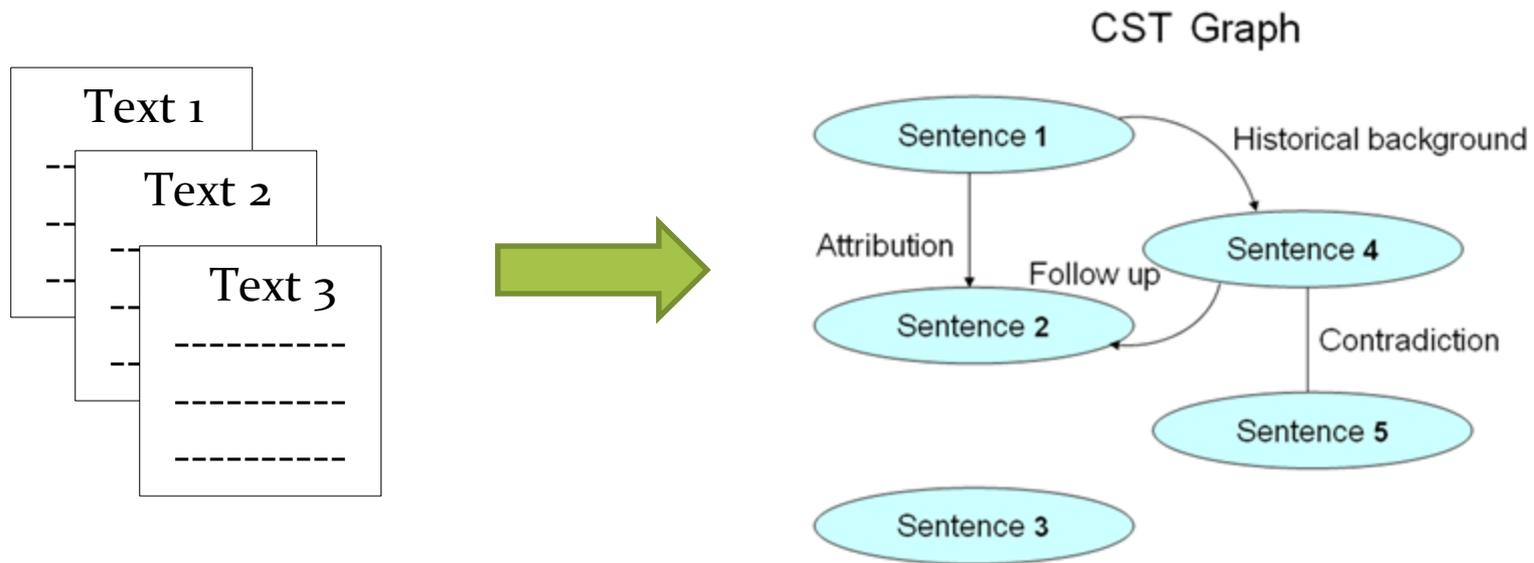
- Tipos e frequências de alinhamento

Tipos de Alinhamento													
1-0	1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12	
2	71	90	67	36	37	13	5	5	1	1	2	1	

- Concordância kappa: 0.831

CROSS-DOCUMENT STRUCTURE THEORY (CST) (RADEV, 2000)

- O primeiro grande desafio com discurso



CROSS-DOCUMENT STRUCTURE THEORY (CST) (RADEV, 2000)

- O primeiro grande desafio com discurso
 - 2 tentativas
 - Versão oficial: quase 1 semestre de anotação
 - 4 anotadores
 - CSTTool: solução caseira

CSTTool

The screenshot shows the CSTTool application window with two tabs: "Text Segmentation" and "CST Structuring". The "CST Structuring" tab is active. The main area contains two text boxes, "Text 1" and "Text 2", each with a file path and "Open" and "Clear" buttons. Below each text box is a list of segmented text items, numbered <1> through <6>. At the bottom, there is a section for selecting relations between segment pairs. It includes a table with columns for "Segment pairs (Text 1 - Text 2)", "CST relation", and "Directionality". A table below shows a relation between segment pairs "2 - 4" with a "Contradiction" relation and "None" directionality. To the right of this table are buttons for "Include", "New CST relation", and "Add". Further right is a text box for "Your name" containing "Priscila". At the bottom, there is a text box for "Relations that you included" containing XML-like relation code: <R SDID="D1_C1_Folha.txt" SSENT="2" TDID="D3_C1_JB.txt" TSENT="4"><RELATION TYPE="Contradiction" JUDGE="Priscila"/></R>. To the right of this text box are buttons for "Open", "Save", and "Clear".

Text Segmentation CST Structuring

Open the texts (already segmented) that you want to analyze and put the relations among their segments using the box in the bottom. Do not forget to identify yourself

Text 1 C:\Documents and Settings\Priscila\Desk Open Text 1 Clear

<1> Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. |

<2> Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade. |

<3> A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto. |

<4> Acidentes aéreos são frequentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética. |

<5> O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes. |

<6> Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 quilômetros. |

Text 2 C:\Documents and Settings\Priscila\Desk Open Text 2 Clear

<1> Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo, matou 17 pessoas na quinta-feira à tarde, informou hoje um porta-voz das Nações Unidas. |

<2> As vítimas do acidente foram 14 passageiros e três membros da tripulação. Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu. |

<3> O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala. "Não houve sobreviventes", disse Okala. |

<4> O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congoleza, a Trasept Congo, também levava uma carga de minerais. |

<5> Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa. |

Select relations and their directionality among the segment pairs that you judge appropriate (you do not need to put relations among all segment pairs)

Segment pairs (Text 1 - Text 2)	CST relation	Directionality
2 - 4	Contradiction	None

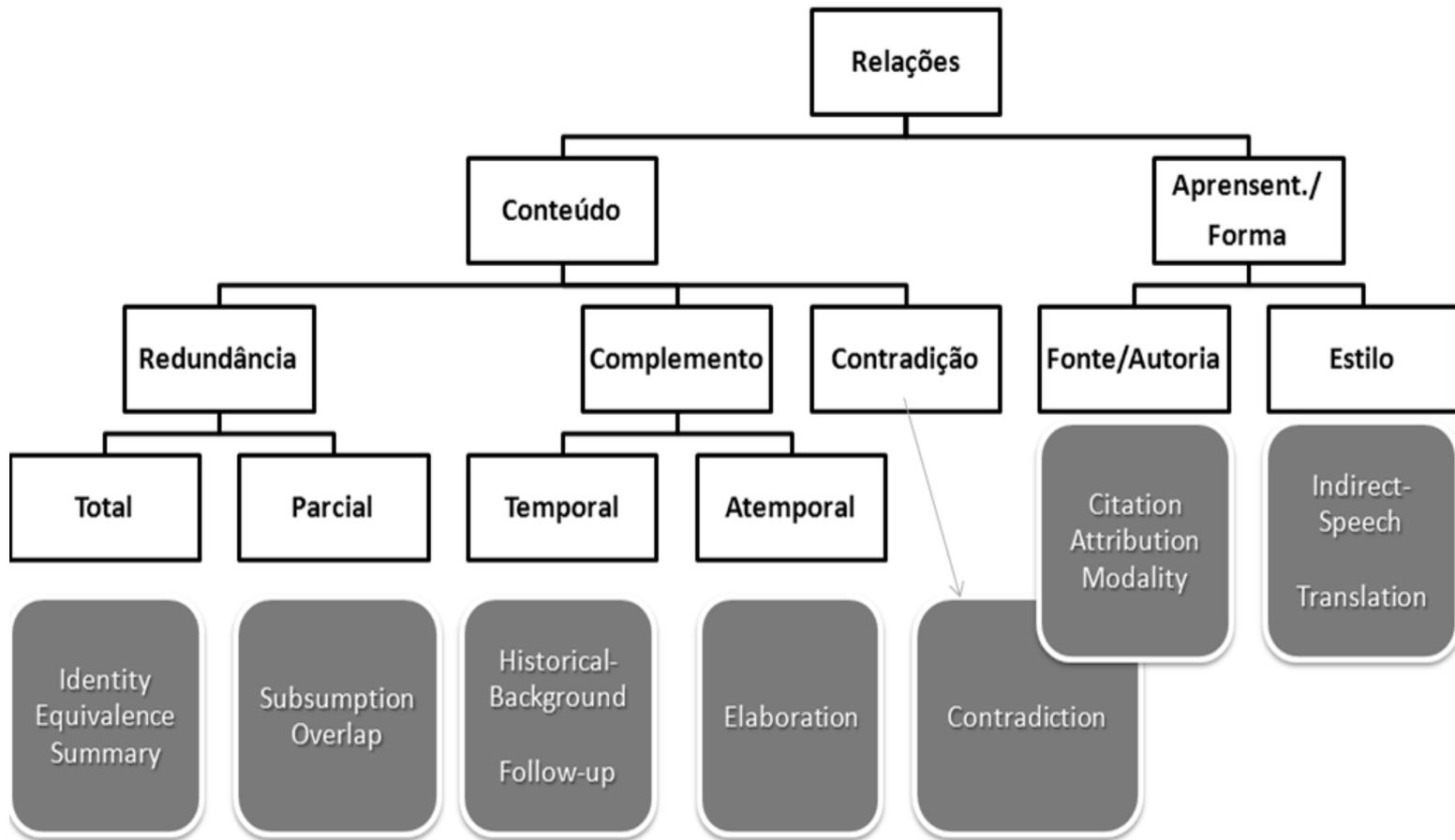
New CST relation Add Your name Priscila

Relations that you included (you may also edit this text box directly if you wish)

```
<R SDID="D1_C1_Folha.txt" SSENT="2" TDID="D3_C1_JB.txt" TSENT="4">
<RELATION TYPE="Contradiction" JUDGE="Priscila"/>
</R>
```

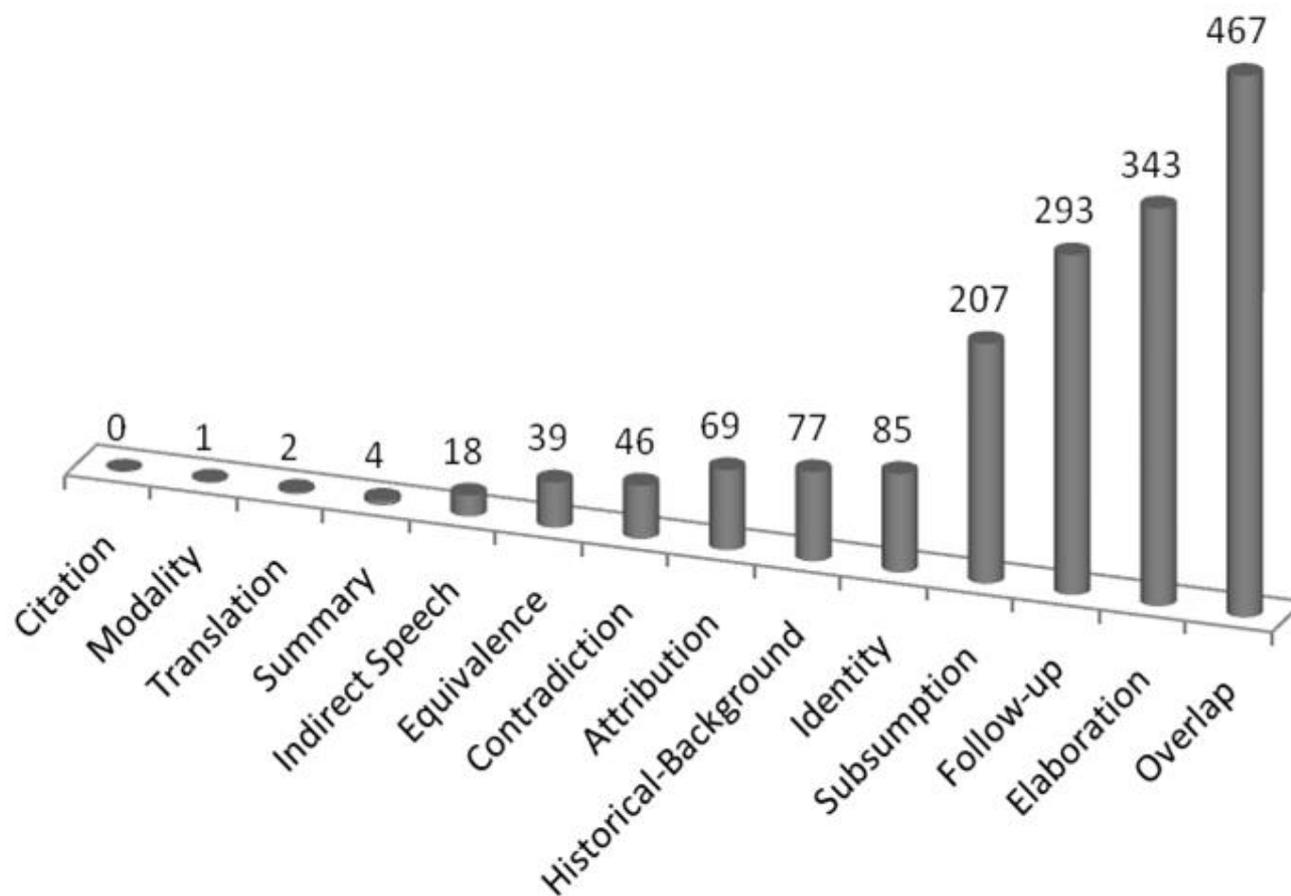
Open Save Clear

TIPOLOGIA DE RELAÇÕES (MAZIERO ET AL., 2014)



CST

- Fenômenos multidocumento



CST

- Confiança na anotação

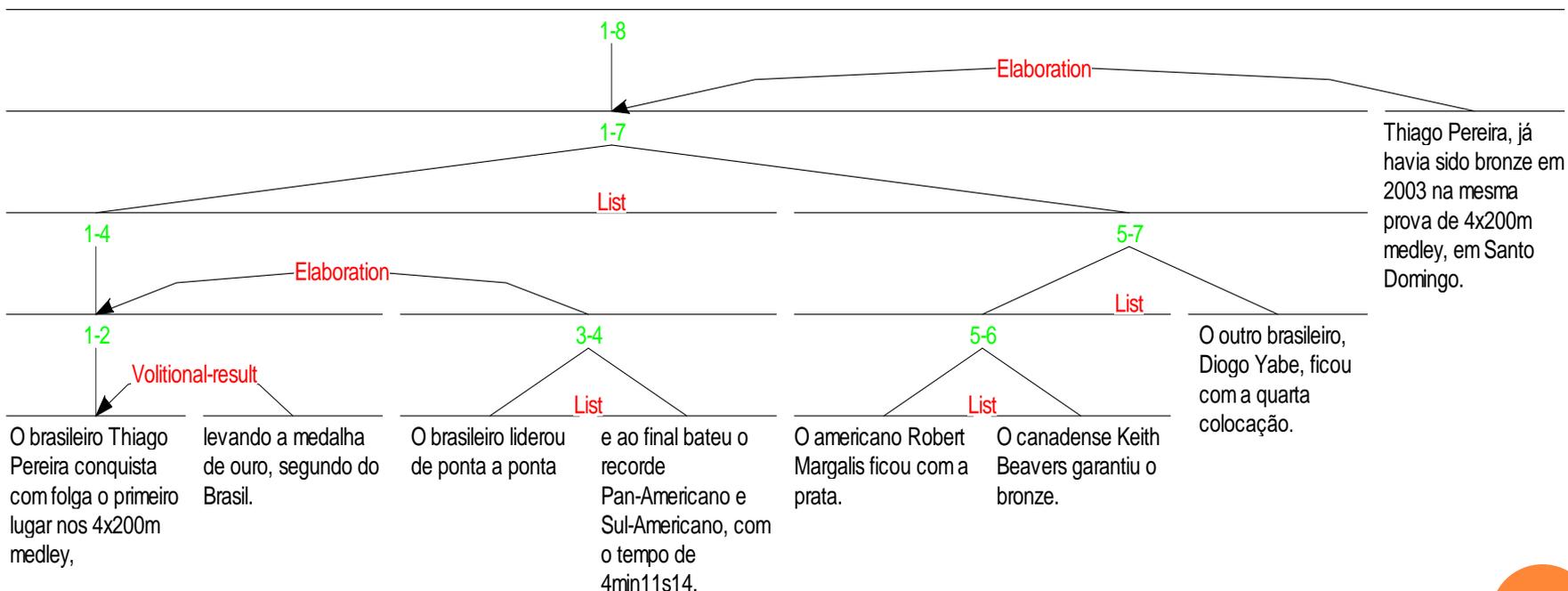
	<i>Kappa</i>	<i>Porcentagem de concordância</i>		
		Total	Parcial	Nula
Relações	0.51	0.54	0.27	0.18
Direcionalidade	0.45	0.58	0.27	0.14
Tipos de relações	0.61	0.70	0.21	0.09

80% de concordância total ou parcial vs. 58% para o inglês

RHETORICAL STRUCTURE THEORY (RST) (MANN E

THOMPSON, 1987)

○ Necessidade do discurso monodocumento



RHETORICAL STRUCTURE THEORY (RST) (MANN & THOMPSON, 1987)

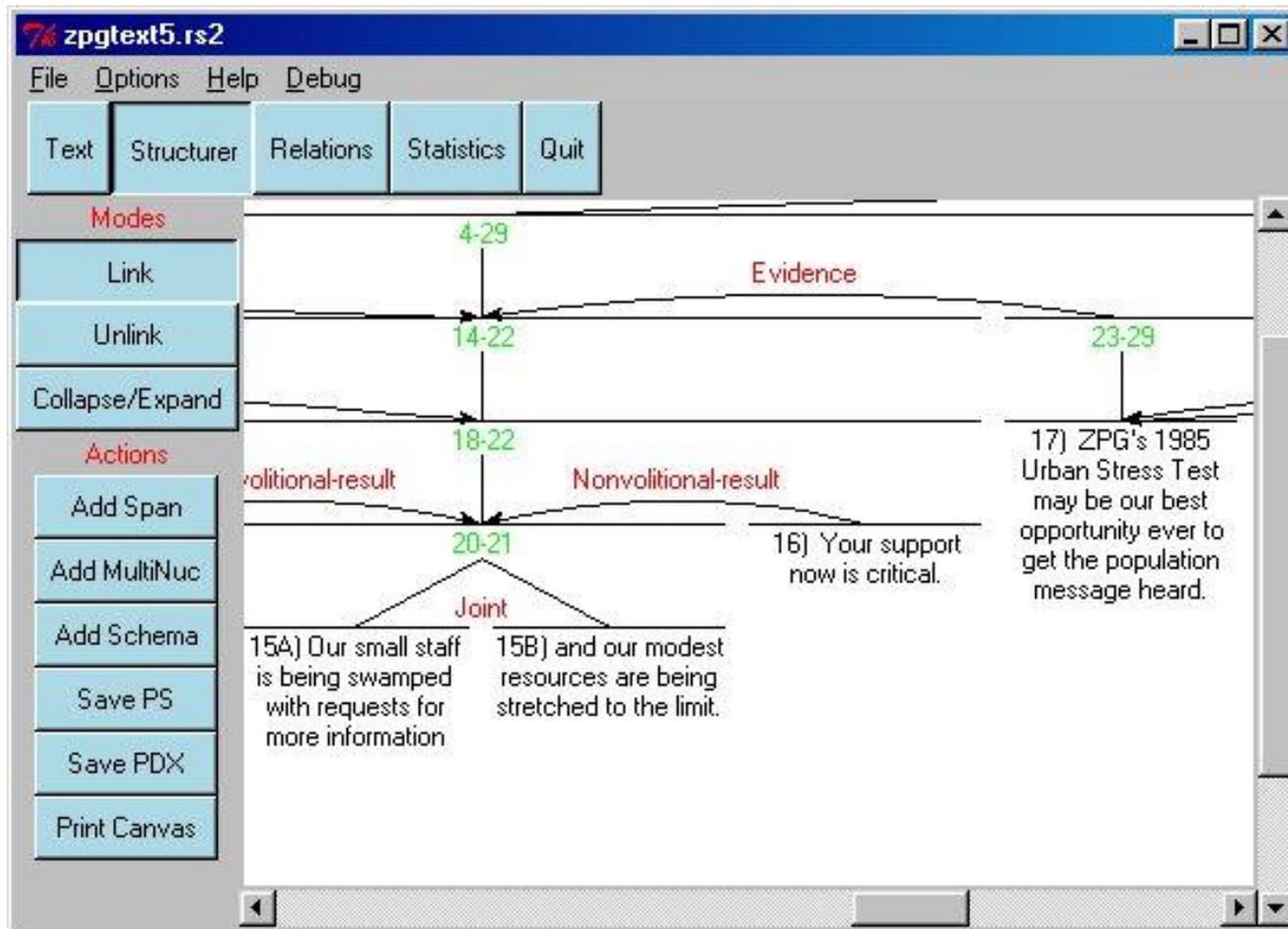
THOMPSON, 1987)

- Anotação de cópús
 - 8 pessoas
 - Quase 1 semestre de anotação
 - RSTTool

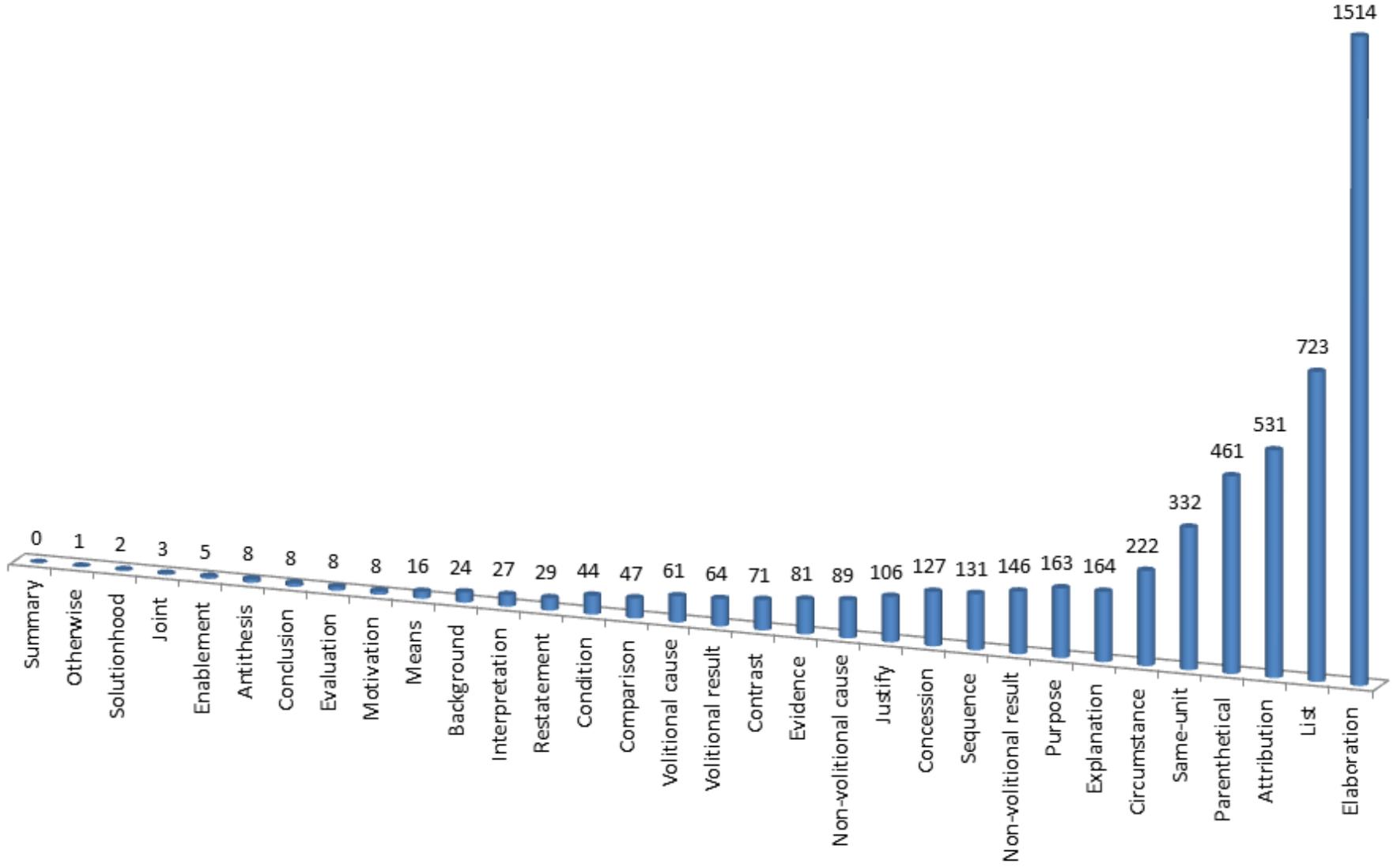
Circumstance	Volitional Cause	Otherwise	Means
Solutionhood	Non-Volitional Cause	Interpretation	List
Elaboration	Volitional Result	Evaluation	Explanation
Background	Non-Volitional Result	Restatement	Comparison
Enablement	Purpose	Summary	Conclusion
Motivation	Antithesis	Sequence	Attribution
Evidence	Concession	Contrast	Parenthetical
Justify	Condition	Joint	Same-Unit

RSTTOOL

- Editor clássico (O'Donnell, 2000)



RST



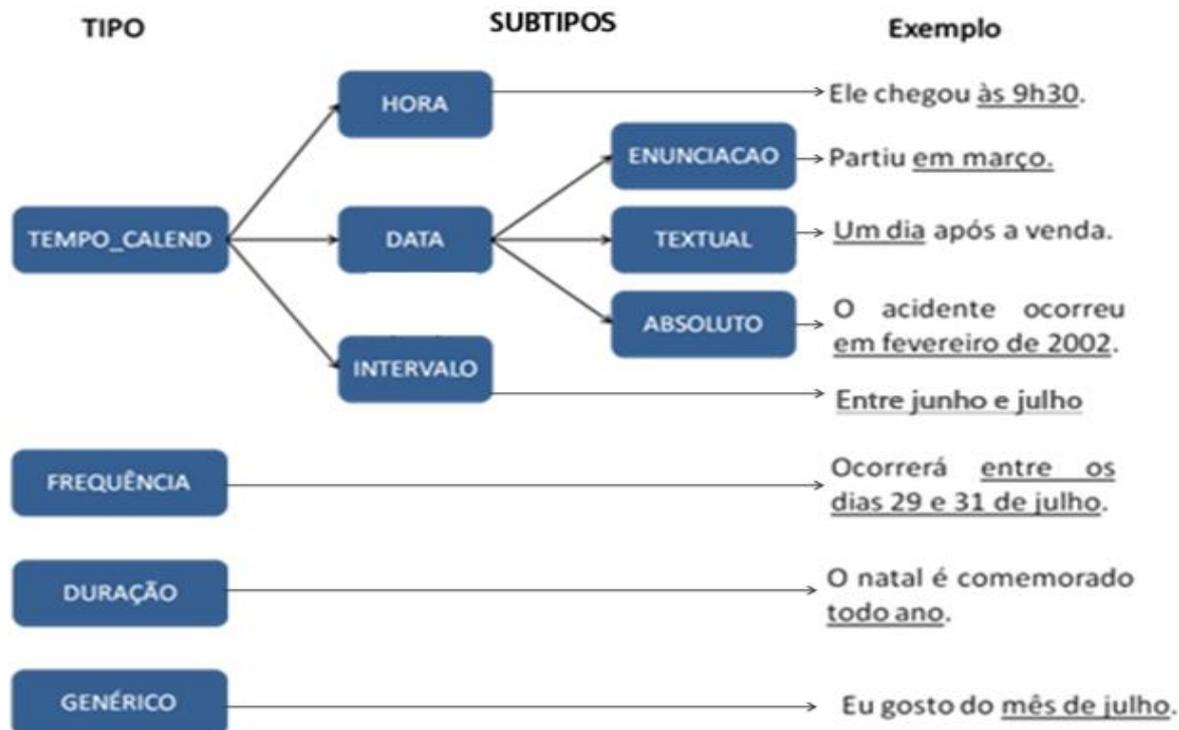
RST

- Concordância com **RSTEval**, com base no método de Marcu (2000)

<i>Evaluated Criteria</i>	<i>Precision</i>	<i>Recall</i>	<i>F-Measure</i>
<i>Simple Textual Segments</i>	<i>0.91</i>	<i>0.91</i>	<i>0.91</i>
<i>Complex Textual Segments</i>	<i>0.78</i>	<i>0.78</i>	<i>0.78</i>
<i>Nuclearity</i>	<i>0.78</i>	<i>0.78</i>	<i>0.78</i>
<i>Relations</i>	<i>0.66</i>	<i>0.66</i>	<i>0.66</i>

ANOTAÇÃO TEMPORAL

- Diretrizes do HAREM (Baptista et al., 2008)
 - 1 único anotador (razões de consistência, tempo e escopo)



ANOTAÇÃO TEMPORAL

○ Exemplo

“O último jogo havia sido <ET ID="03" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="200510--T----E--LM-"> em outubro de 2005 </ET>, na vitória por 3 a 0 sobre a Venezuela (...).”

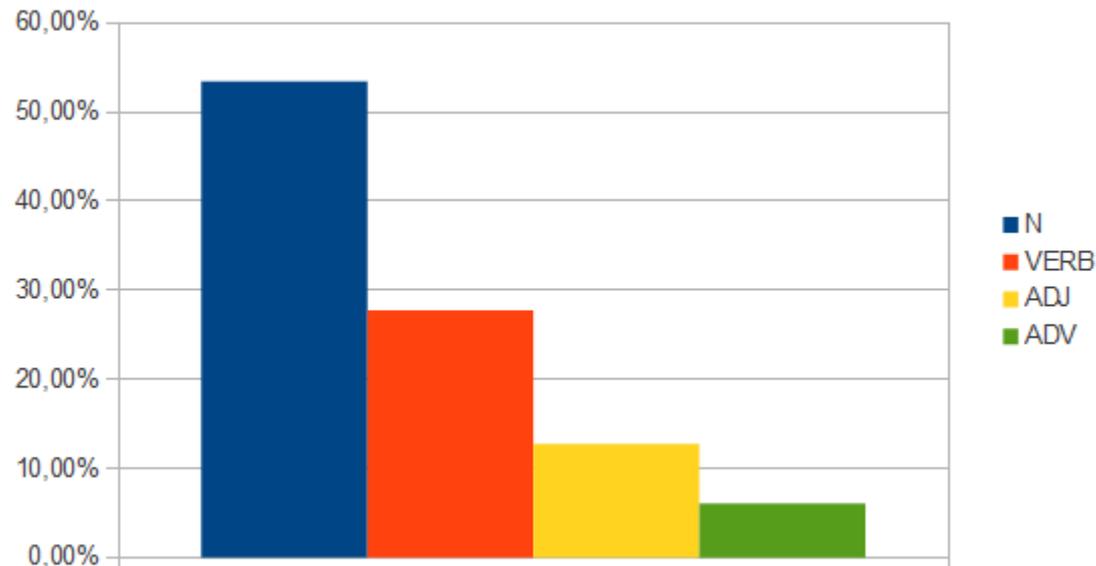
Frequência no corpus

Tipo de ET	Ocorrência
Absoluto	121
Duração	54
Frequência	6
Intervalo	76
Hora	122

Tipo de ET	Ocorrência
Enunciação	522
Textual	28
Genérico	1
Sem classificação	3
Cabeçalhos	140

SENTIDO

- Indexação dos substantivos comuns e verbos a synsets da wordnet de Princeton

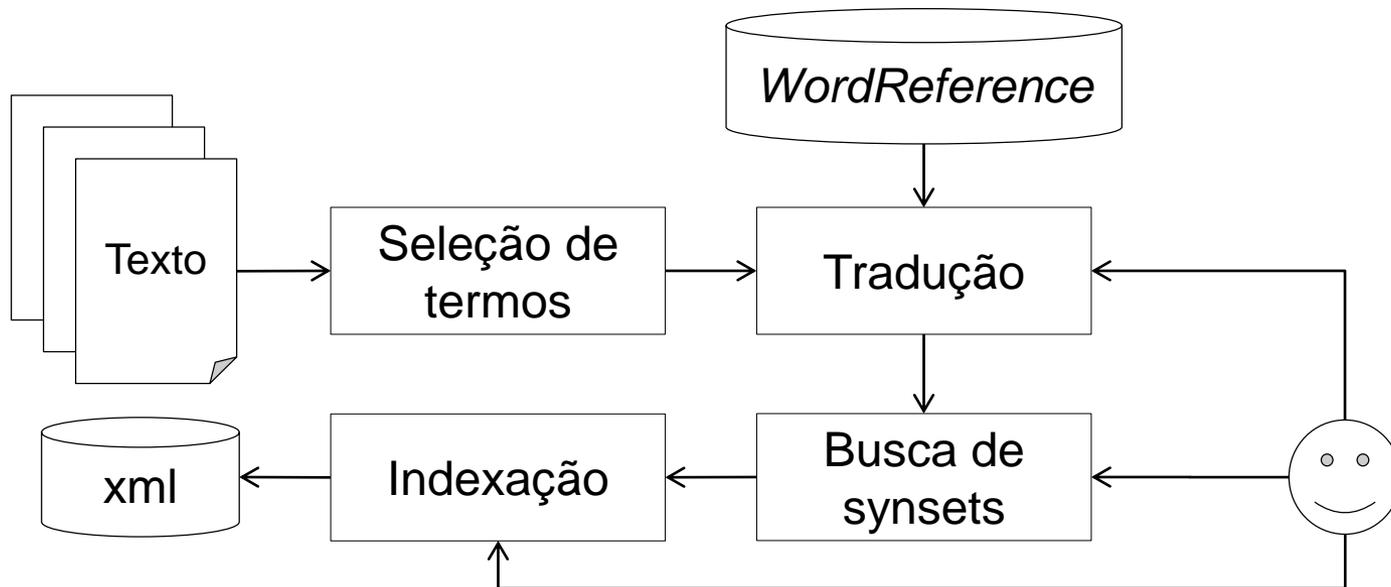


Nóbrega, F.A.A. and Pardo, T.A.S. (2014). General Purpose Word Sense Disambiguation Methods for Nouns in Portuguese. In the Proceedings of the PROPOR 2014 PhD and MSc/MA Dissertation Contest / 11st International Conference on Computational Processing of Portuguese - PROPOR (LNAI 8775), pp. 94-101. October 6-9. São Carlos-SP/Brazil.

Sobrevilla Cabezudo, M.A.; Maziero, E.G.; Souza, J.W.C.; Dias, M.S.; Cardoso, P.C.F.; Balage Filho, P.P.; Agostini, V.; Nóbrega, F.A.A.; Barros, C.D.; Di Felippo, A.; Pardo, T.A.S. (2015). Anotação de Sentidos de Verbos em Textos Jornalísticos do Corpus CSTNews. Revista de Estudos da Linguagem - RELIN, Vol. 23, N. 3, pp. 797-832.

SENTIDO

- Ferramenta de auxílio à anotação
 - 10 anotadores
 - Cerca de 3 meses de anotação



NASP

The screenshot displays the NASP application window with a menu bar (Arquivo, Anotação, Ajuda) and a 'Visualizador de texto' (Text Viewer) section. The text viewer shows a paragraph with words like 'avião' and 'passageiros' highlighted in red boxes. Below the text viewer, there are two main panels: '1 - Escolha a tradução' (Choose the translation) and '2 - Escolha o Synset' (Choose the Synset). Panel 1 shows a list of possible translations for 'avião', including 'plane' and 'airplane'. Panel 2 shows a list of synsets for 'avião', including 'airplane', 'sheet', and 'carpenter's plane'. At the bottom, there are buttons for 'Nova tradução' (New translation), 'Adicionar nova tradução' (Add new translation), 'Selecionar Synset' (Select Synset), and 'Remover anotação' (Remove annotation).

Visualizador de texto

Arquivo Anotação Ajuda

Texto 1 Texto 2 Texto 3

ram após a queda de um **avião** de **passageiros** na F
ONU , o **avião** , de fabricação russa , estava tentanc
uma montanha e caiu , em chamas , sobre uma flores
entes no Congo , onde 51 **companhias** privadas oper
do pela Air Traset , levava 14 **passageiros** e três trip
ineira de Lugushwa em direção a Bukavu , numa dist
amente para transporte na República Democrática d

1 - Escolha a tradução
Possíveis traduções:

plane
airplane

Nova tradução: Adicionar nova tradução

2 - Escolha o Synset

Ver Hiperônimos Ver Hipônimos

Synsents:

1 -- {airplane, aeroplane, plane}
an aircraft that has a fixed wing and is powered by p
● Ex. 1: "the flight was delayed due to trouble w

2 -- {plane, sheet}
(mathematics) an unbounded two-dimensional shape
● Ex. 1: "we will refer to the plane of the graph a
● Ex. 2: "any line joining two points on a plane li

3 -- {plane}
a level of existence or development
● Ex. 1: "he lived on a worldly plane"

4 -- {plane, planer, planing machine}
a power tool for smoothing or shaping wood

5 -- {plane, carpenter's plane, woodworking plane}
a carpenter's hand tool with an adjustable blade for
● Ex. 1: "the cabinetmaker used a plane for the f

Selecionar Synset Remover anotação

SENTIDO

- Concordância

Kappa

	<u>Substantivos</u>	<u>Verbos</u>
Tradução	0,853	0,648
Synset	0,729	0,509
Tradução+Synset	0,697	0,474

SUBTÓPICOS

- Tópico principal do texto desenvolvido por subtópicos (Hearst, 1997)
 - 14 anotadores
 - Interface de anotação “sofisticada”: notepad ;-)

EXEMPLO

[S1] A plane crash in Bukavu, in the Eastern Democratic Republic of Congo, killed 17 people on Thursday afternoon, said the spokesman of the United Nations.

[S2] The victims of the accident were 14 passengers and three crew members.

[S3] All died when the plane, hampered by the bad weather, failed to reach the runway and crashed in a forest that was 15 kilometers from the airport in Bukavu.

[S4] The plane exploded and caught fire, said the UN spokesman in Kinshasa, Jean-Tobias Okala.

[S5] “There were no survivors”, said Okala.

<subtopic: plane crash in the Congo>

[S6] The spokesman said the plane, a Soviet Antonov-28 and Ukrainian manufacturing and ownership of the Trasept Congo, a Congolese company, also took a mineral load.

<subtopic: details about the plane >

[S7] According to airport sources, the crew members were Russian.

<subtopic: details about the flight crew>



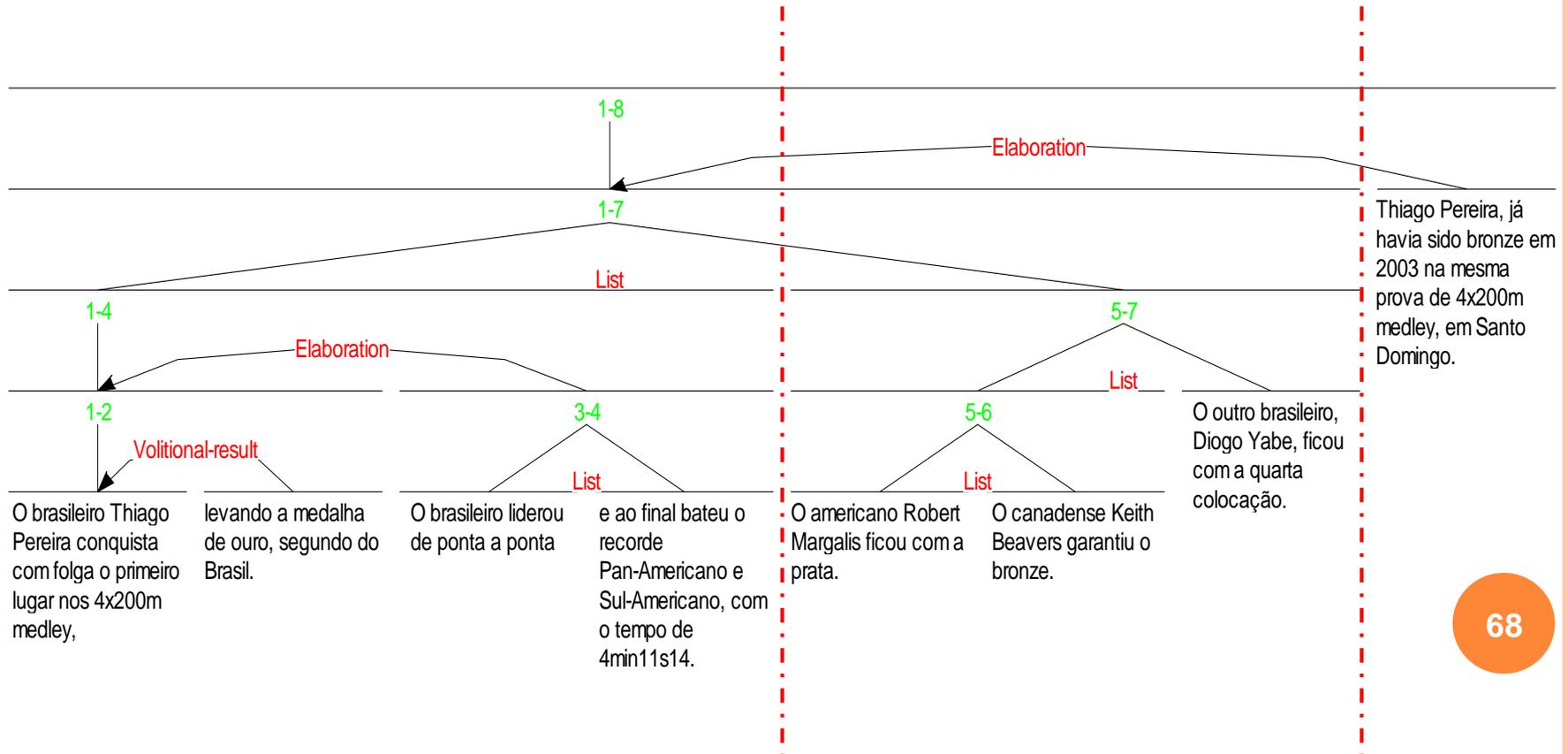
SUBTÓPICOS

- Concordância

Day	Groups	Number of annotators	Texts per group	Kappa
1	A	6	10	0.656
	B	7		0.566
2	A	5	10	0.458
	B	5		0.447
3	A	7	10	0.515
	B	5		0.638
4	A	5	10	0.544
	B	7		0.562
5	A	5	10	0.643
	B	5		0.528
6	A	5	12	0.570
	B	5	13	0.549
7	A	5	15	0.611
Average				0.560

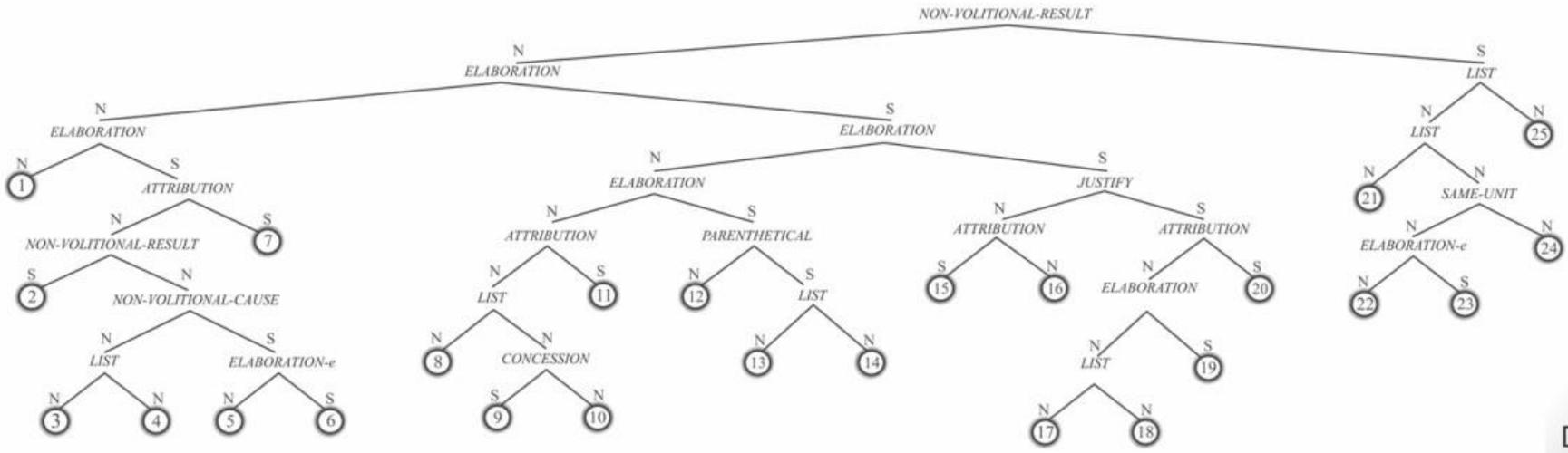
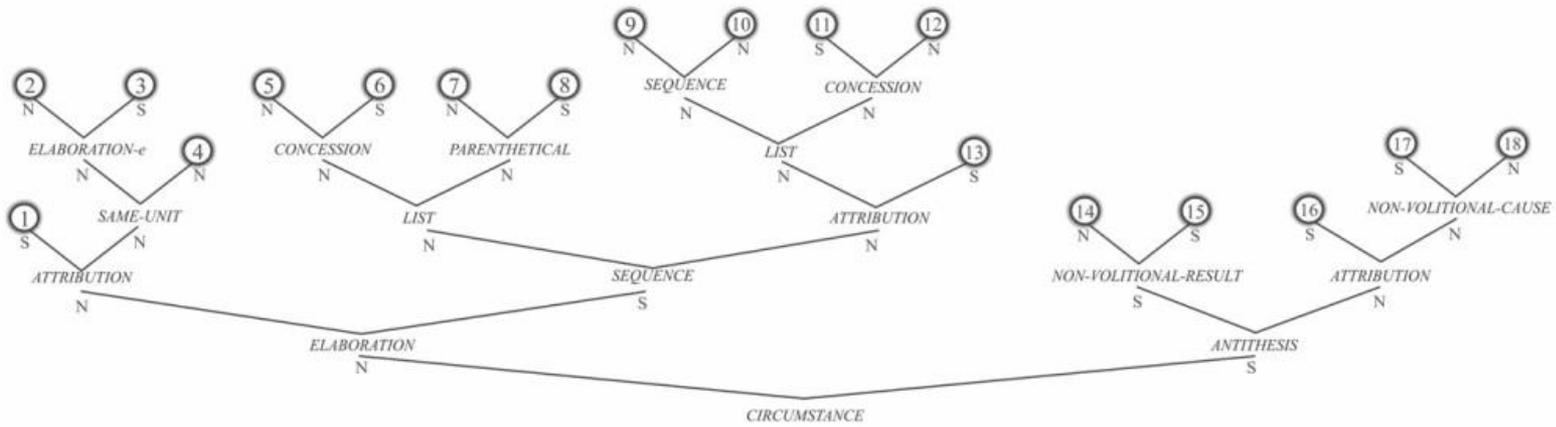
RST & SUBTÓPICOS

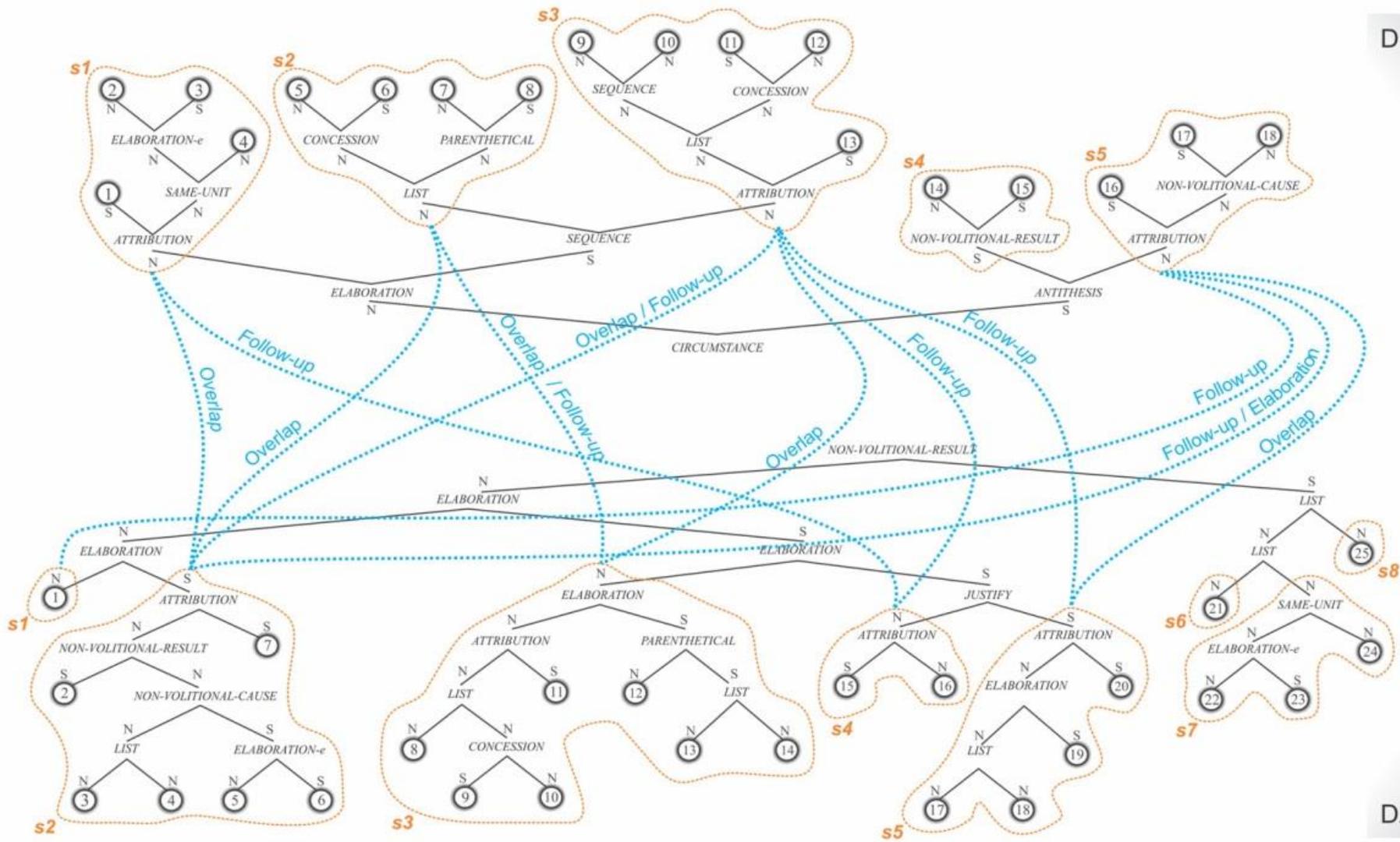
○ Correspondência entre subtópicos e estruturação discursiva

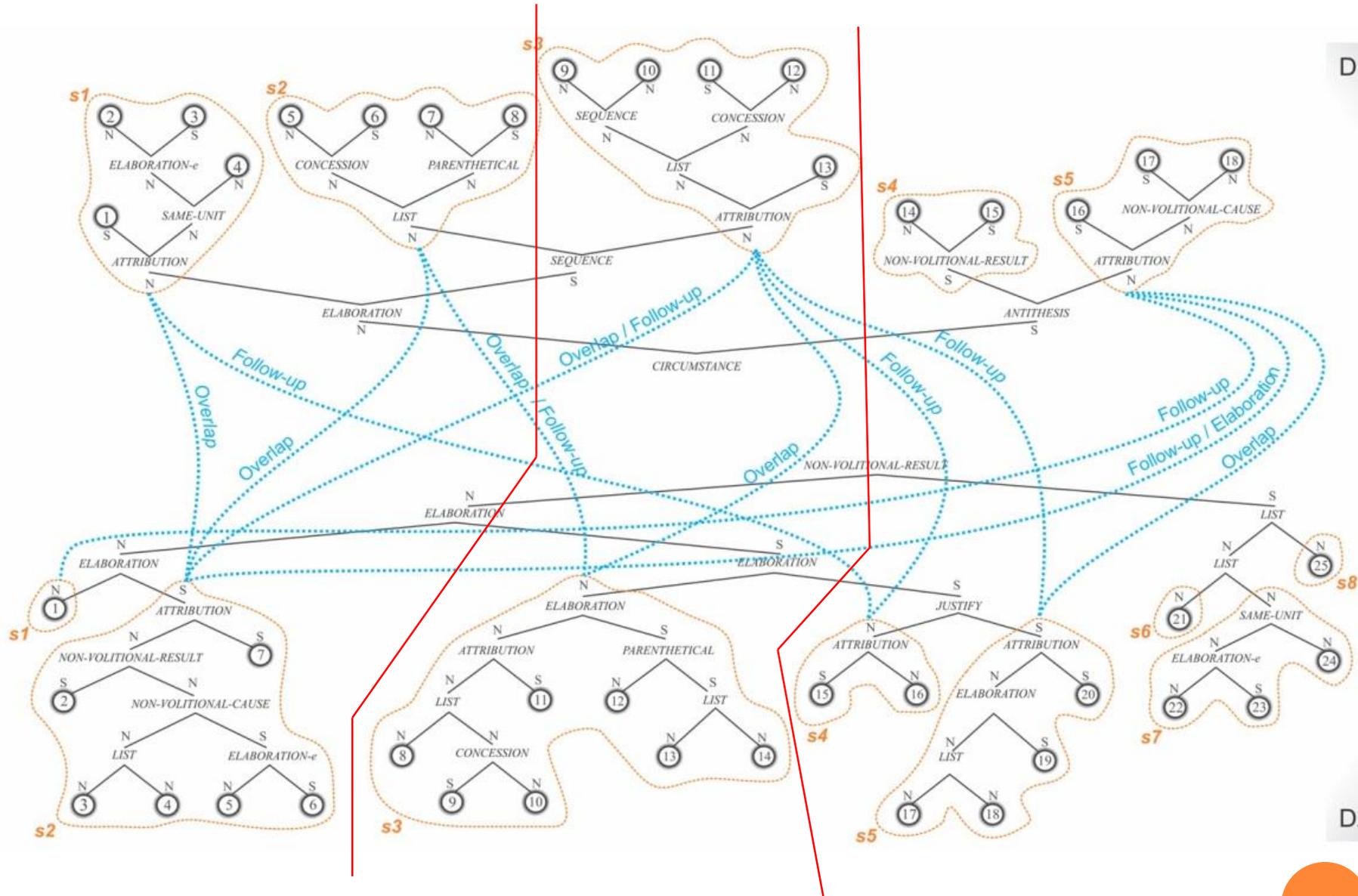


POR QUE PARAR NISSO?

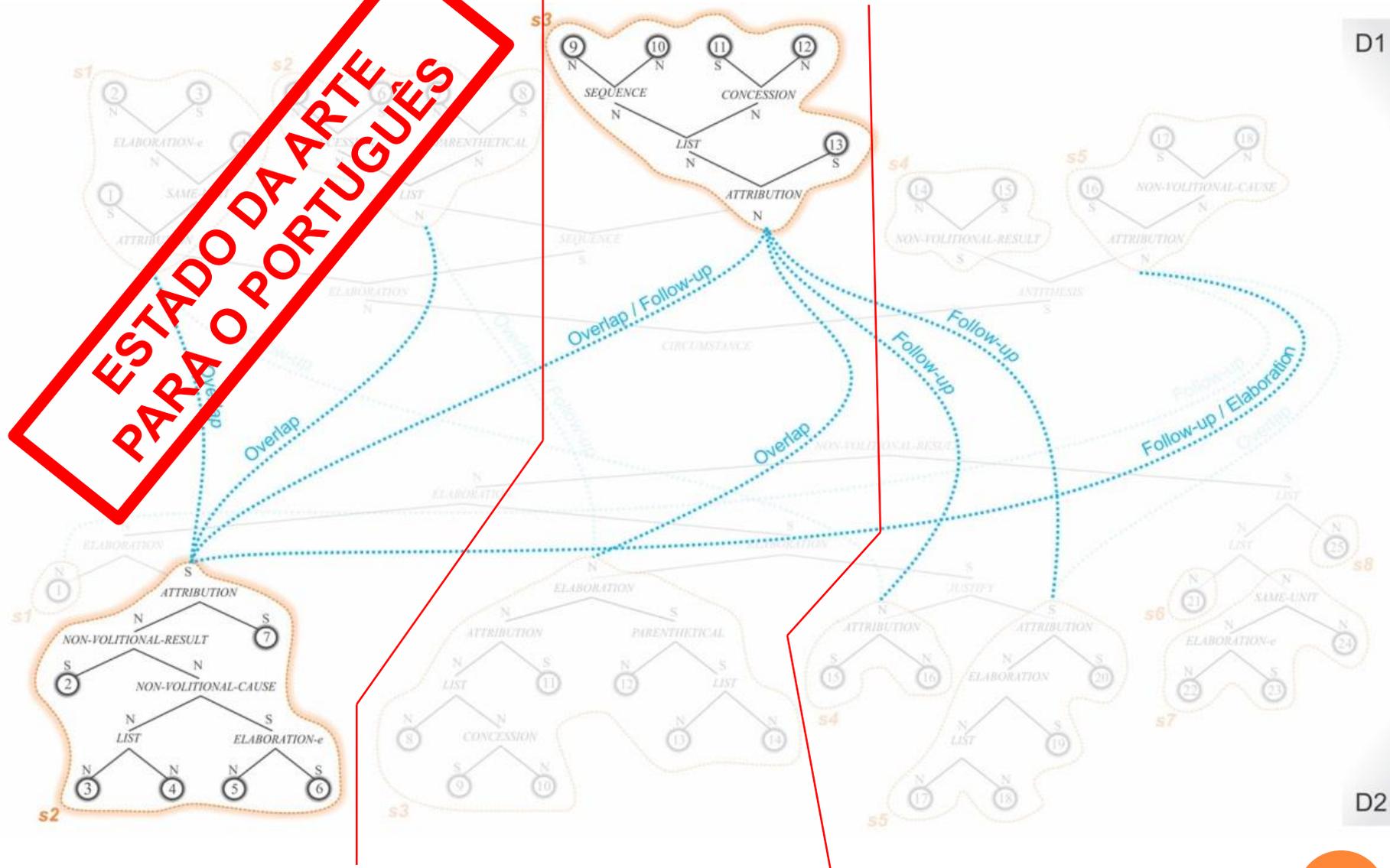
- O grande diferencial do CSTNews: correlação entre níveis até então inexplorada







**ESTADO DA ARTE
PARA O PORTUGUÊS**



D1

D2

CORREFERÊNCIAS

- O outro nível discursivo “clássico”
 - Relações discursivas: coerência global
 - Correferências: coerência local

- Vários ensaios para a anotação
 - Oportunidade surgiu no IberEval 2017, em uma tarefa de anotação colaborativa

Pardo, T.A.S.; Baptista, J.; Duran, M.S.; Nunes, M.G.V.; Nóbrega, F.A.A.; Aluísio, S.M.; Di Felippo, A.; Seno, E.R.M.; Silva, R.R.; Anchiêta, R.T.; Brum, H.B.; Dias, M.S.; Martins, R.S.O.; Maziero, E.G.; Souza, J.W.C.; Vargas, F.A. (2017). The Coreference Annotation of the CSTNews Corpus. In the Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval), pp. 102-112. September, 19. Murcia/Spain.

CORREFERÊNCIAS

- Exemplo monodocumento

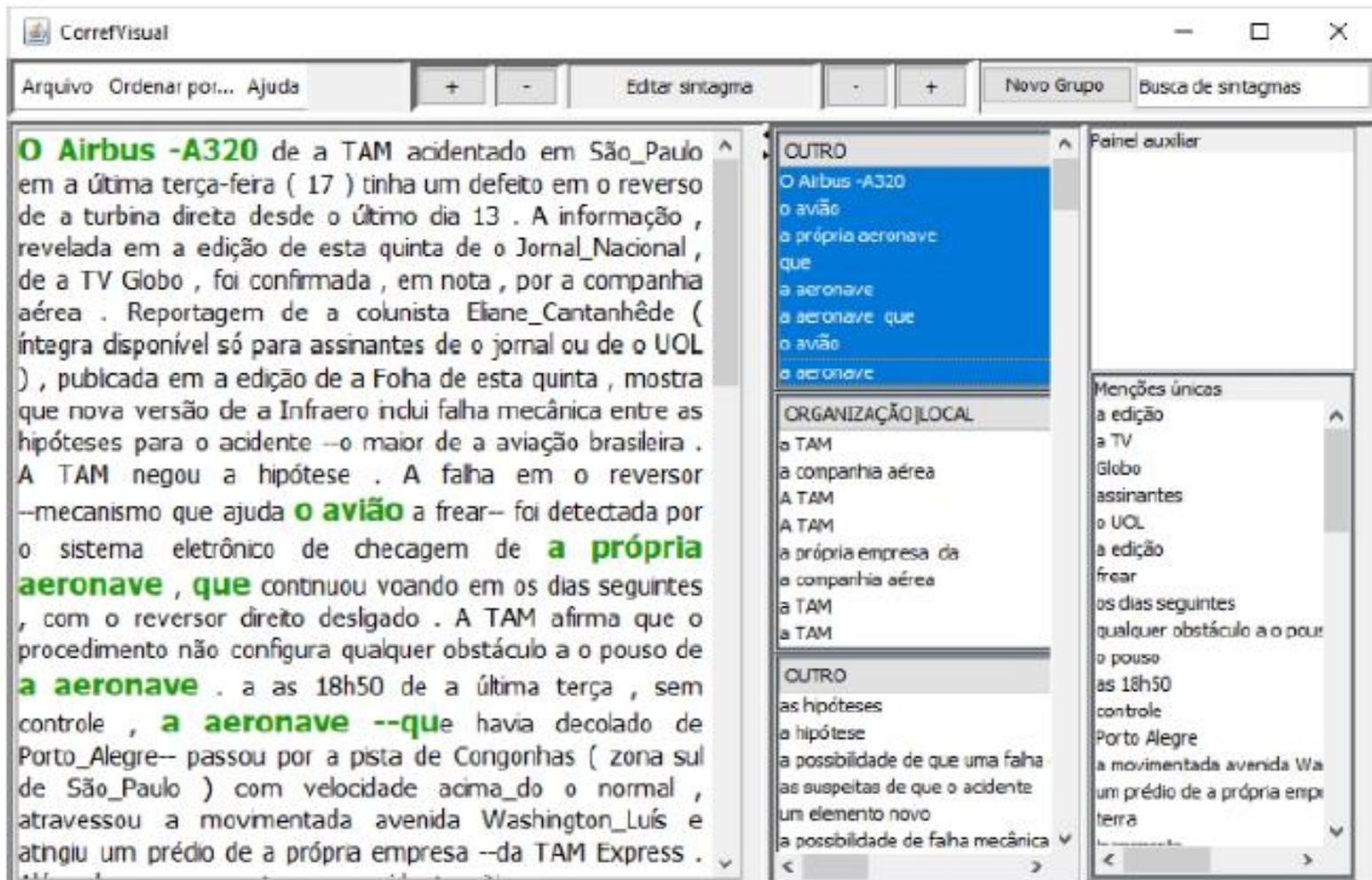
At least 17 people died after the crash of a passenger plane in the Democratic Republic of Congo. According to an ONU spokeswoman, the airplane was trying to land in the Bukavu airport in the midst of a storm. It failed to reach the runway and fell in a forest 15 kilometers away from the airport.

- Mas nosso problema é multidocumento!

- 5 times de anotação, com 2 a 3 pessoas por time e um pesquisador mais sênior
- 3 semanas de anotação, 1 hora diária

CORREFVISUAL

- Desafios da definição do fenômeno, passando pelo pré-processamento automático até o editor de anotação



CORREFERÊNCIA

- Concordância

Teams	Kappa
1	0.50
2	0.48
3	0.55
4	0.64
5	0.57
<i>Average</i>	<i>0.54</i>

ASPECTOS INFORMATIVOS

- Em um sumário, identificar “o que” aconteceu, “quando”, “onde”, etc.
 - 4 times, com 3 anotadores cada

Summary in Portuguese (original)

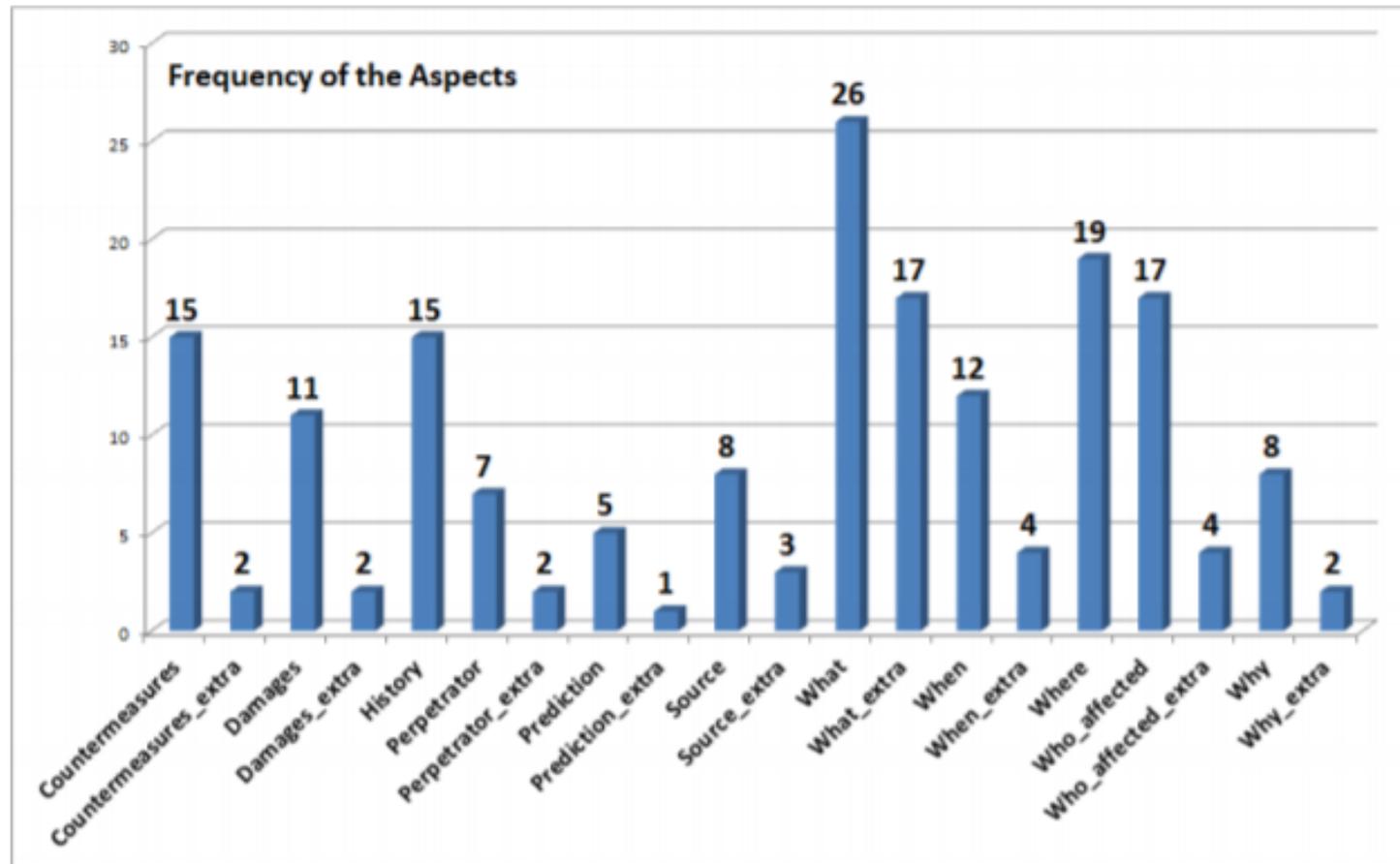
[17 pessoas morreram]WHO_AFFECTED [após a queda de um avião]WHAT [na República Democrática do Congo.]WHERE [14 dessas vítimas eram passageiros e três membros da tripulação, todos de nacionalidade russa.]WHO_AFFECTED [Nenhuma vítima sobreviveu.]DAMAGES_EXTRA

[O avião saiu de Lugushwa a Bukavu e caiu sobre uma floresta]WHERE [após se chocar com uma montanha, prejudicado pelo mau tempo.]WHY

[O avião também levava cargas e minerais.]WHAT_EXTRA

ASPECTOS INFORMATIVOS

- Sumários de “mundo”



ASPECTOS INFORMATIVOS

- Exemplo de padrões “aprendidos” por tópico

	Accidents	Attacks	Legal and Political Decisions	Natural Disasters
For all summaries				
<i>In common</i>	What, Where, Who_affected, Why	What, Where, Who_affected	What, Perpetrator	What, Where, Who_affected, Countermeasures, Damages
<i>In the 1st paragraph</i>	What, Where, Who_affected	What, Where	What, Perpetrator	What, Where
<i>Partial ordering</i>	What < Where Who_affected, What, Where < Why	What < Where	---	What < Where What, Where < Countermeasures, Damages
For the majority of summaries				
<i>In common</i>	---	When, Perpetrator, Why, History	History	When, Prediction
<i>In the 1st paragraph</i>	---	Perpetrator, When	---	Who_affected, When, Damages
<i>Partial ordering</i>	---	---	Who_affected, What, Perpetrator < History	What, Where < Who_affected

OUTRAS ANOTAÇÕES

- Ontologias derivadas das anotações de sentido
 - Recortes sobre a Wordnet de Princeton
- Anotação automática de morfossintaxe e sintaxe pelo parser PALAVRAS (Bick, 2000)
- Ideia principal dos textos-fonte
- Etc.

CÓRPUS CSTNEWS

- **Diversas camadas de anotação, completamente disponível**, servindo de base para muitos trabalhos

Análise morfofossintática e sintática dos textos-fonte

Aspectos nos sumários multidocumento

Alinhamento textos-sumários multidocumento

Subtópicos nos textos-fonte

Sentidos dos substantivos mais frequentes e verbos dos textos-fonte

Anotação CST dos textos-fonte

Anotação RST dos textos-fonte

Expressões temporais nos textos-fonte

Sumários automáticos multidocumento, ordenados e não ordenados

Sumários humanos mono e multidocumento

LIÇÕES APRENDIDAS

- Anotação de corpus é uma ciência (Hovy e Lavid, 2010)
 - Processo sistemático é necessário
 - Humano: normalmente o mais importante, mas traz grandes desafios!
 - Motivação e grande contribuição
 - Grande aprendizado
 - Afinidade com tarefa, atenção com a língua
 - Projeto da tarefa: tempo disponível, cansaço, erro
 - Da subjetividade à “mecanização” da tarefa
 - Pensar só no Aprendizado de Máquina é um erro que pode custar muito caro!
 - Dilema da pesquisa básica vs. aplicada

Caso real

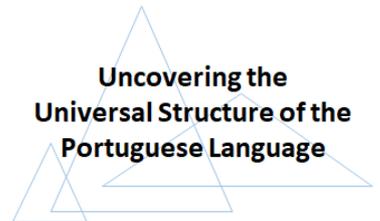
O PROJETO POETISA

POeTiSA

*POrtuguese processing – Towards
Syntactic Analysis and parsing*



Universal Dependencies



POeTiSA

POrtuguese processing - Towards Syntactic Analysis and parsing



Uncovering the
Universal Structure of the
Portuguese Language

POeTiSA is a long term project that aims at growing syntax-based resources and developing related tools and applications for Brazilian Portuguese language, looking to achieve world state-of-the-art results in this area. On the resource side, we focus on the production of a large and comprehensive multi-genre corpus of [Universal Dependencies](#)-based part of speech and syntactically annotated texts, including mainly news texts and user-generated content (tweets and online comments). Regarding the tools, we aim to investigate recent neural and distributional-based methods for training robust parsing models for Portuguese. The project also envisions the production of applications on opinion mining and sentiment analysis tasks that may benefit from syntactic knowledge, as opinion summarization, helpfulness prediction, aspect identification, deception detection and emotion classification.

This project is part of the [Natural Language Processing initiative \(NLP2\)](#) of the [Center for Artificial Intelligence \(C4AI\)](#) of the University of São Paulo, sponsored by IBM and FAPESP (grant #2019/07665-4). The center is part of the [FAPESP Engineering Research Centers Program](#) and is committed to state-of-the-art research in Artificial Intelligence, exploring both foundational issues and applied research.

POeTiSA

- <https://sites.google.com/icmc.usp.br/poetisa>
- A partir de 2020, com até 10 anos de duração



Objetivos

- Treebank multigênero de tamanho significativo para o português
 - Filiado ao modelo *Universal Dependencies* (Nivre, 2015; Nivre et al., 2020)
- Sistemas de parsing completo e parcial para o português, com capacidade de análise multigênero
 - Avaliação e aprimoramento de parsers atuais (Straka, 2018; Zilio et al., 2018)
 - Investigação e desenvolvimento de novos métodos para o português, principalmente com base em abordagens neurais e distribucionais, com resultados do estado da arte
- Aplicação principal relacionada: análise de sentimentos
 - Tarefas que podem se beneficiar da sintaxe: extração de aspectos, detecção de conteúdo enganoso, classificação de utilidade de comentários, classificação de emoções, etc.
 - Avanços metodológicos: sumarização de opiniões, detecção de transtornos psicológicos, etc.



	Número de sentenças	Tamanho médio das sentenças	Número de tokens	Número de types
Folha-Kaggle	3.556.700	21,30	75.818.329	422.228
MAC-MORPHO	43.519	17,40	757.574	49.661
B2W-reviews1	238.567	12,60	2.995.379	55.919
Tweets_stocks	7.281	10,80	78.791	9.286
Comentários de livros	422	19,00	8.058	2.485
Total	3.846.489	16,22	79.658.131	539.579



PorTinari
Portuguese Treebank

folha-kaggle
mac-morpho (Fonseca et al., 2015)
b2w-reviews1 (Real et al., 2019)
Seleção de córpus para treebank
comentários de livros
tweets de bolsa de valores (Silva et al., 2020)



Guerra,
1954

Paz,
1952

Seleção de anotadores e
anotação paralela *issue-based*

Pré-processamento, anotação automática, editor
de anotação

folha-kaggle

mac-morpho
(Fonseca et al., 2015)

Seleção de cópús

para treebank

tweets de

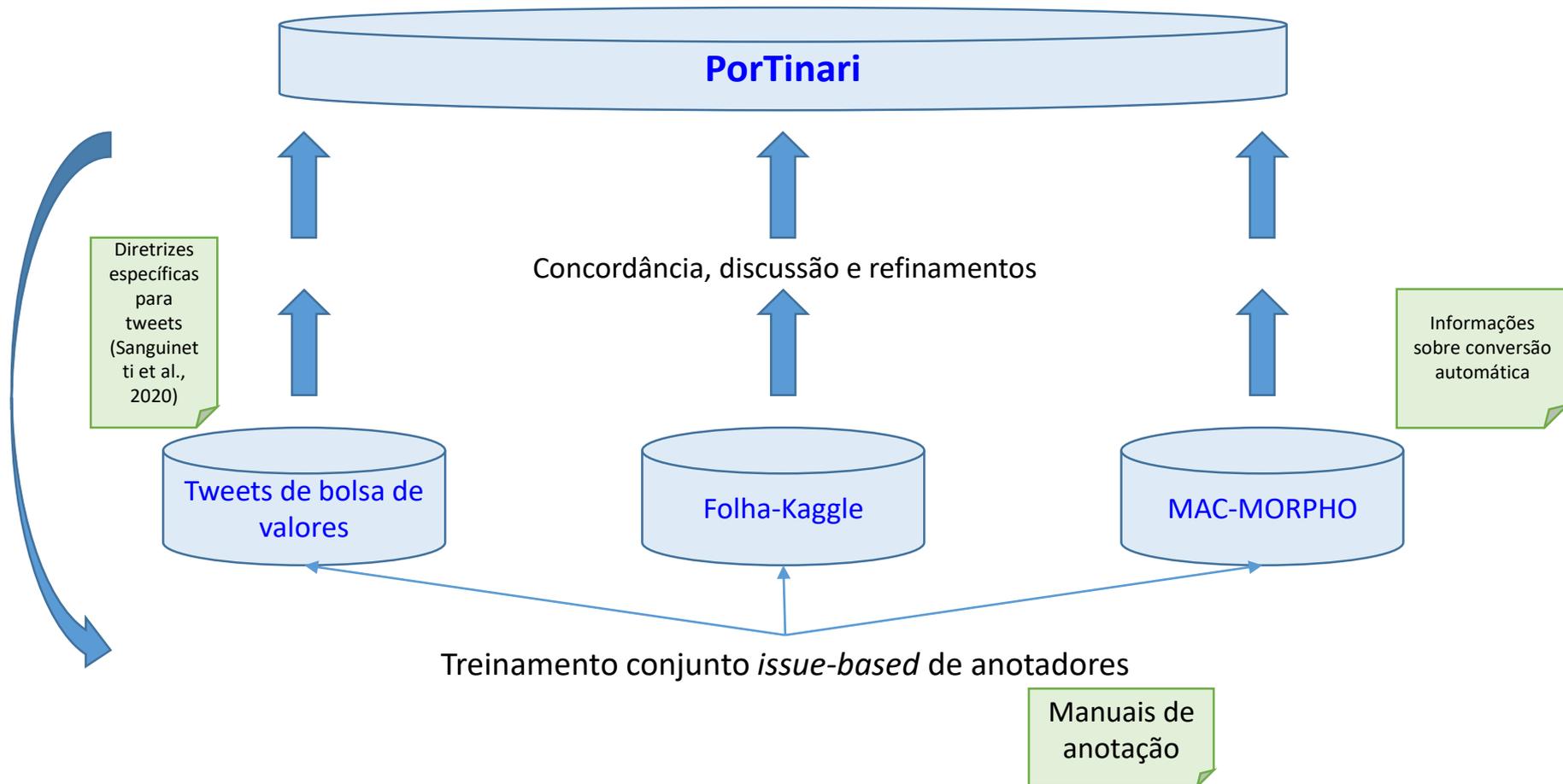
b2w-reviews1
(Real et al., 2019)

bolsa de valores
(Silva et al., 2020)

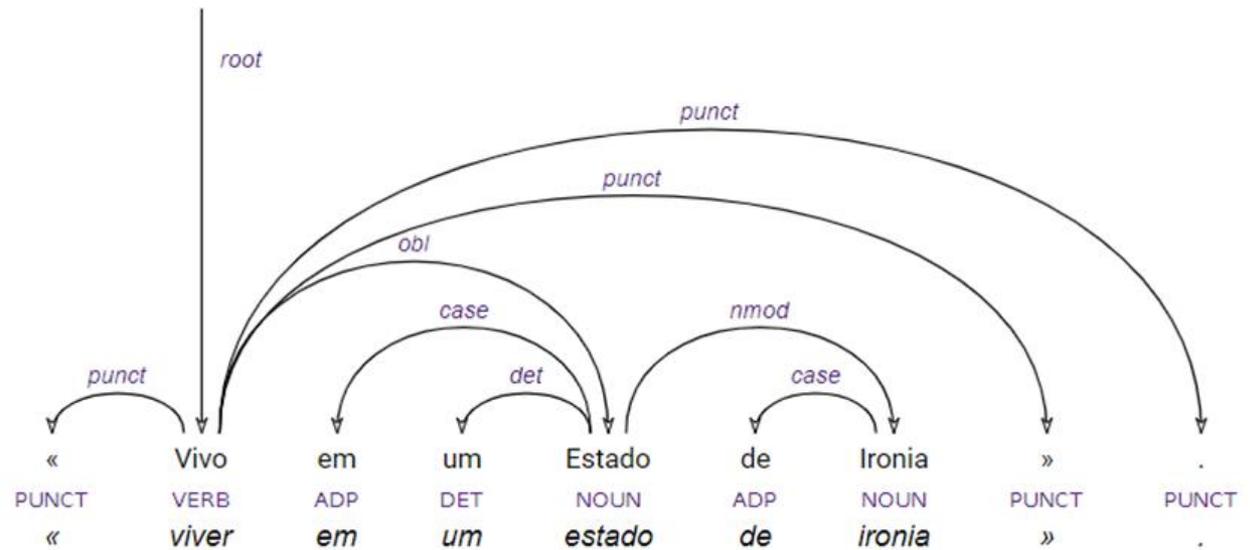
comentários
de livros

Protocolo de anotação, manual de anotação de POS tags,
manual de anotação de relações sintáticas

Anotação paralela *issue-based*: o início



Trebank Annotation for Human Beings



Análise de sentimentos: extração de aspectos, detecção de conteúdo enganoso, classificação de utilidade, classificação de emoções e detecção de transtornos psicológicos

Construção de recursos: léxicos especializados, gramáticas locais

Avaliação de parsers existentes e proposta de regras de pós-edição para correção de erros

Seleção de anotadores e anotação paralela *issue-based*

Pré-processamento, anotação automática, editor de anotação

Difusão

folha-kaggle

mac-morpho
(Fonseca et al., 2015)

Seleção de cópulas para treebank

b2w-reviews1
(Real et al., 2019)

tweets de bolsa de valores
(Silva et al., 2020)

comentários de livros

Protocolo de anotação, manual de anotação de POS tags, manual de anotação de relações sintáticas

Chunking e parsing parcial

Desenvolvimento de taggers e parsers

Sintaxe e estilo de escrita para identificação de autoria

Tarefas da semana

Leitura

- Sabou, M.; Bontcheva, K.; Derczynski, L.; Scharl, A. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In the Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), pp. 859-866.
- No e-Disciplinas

Provinha 7 disponível à tarde no e-Disciplinas