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

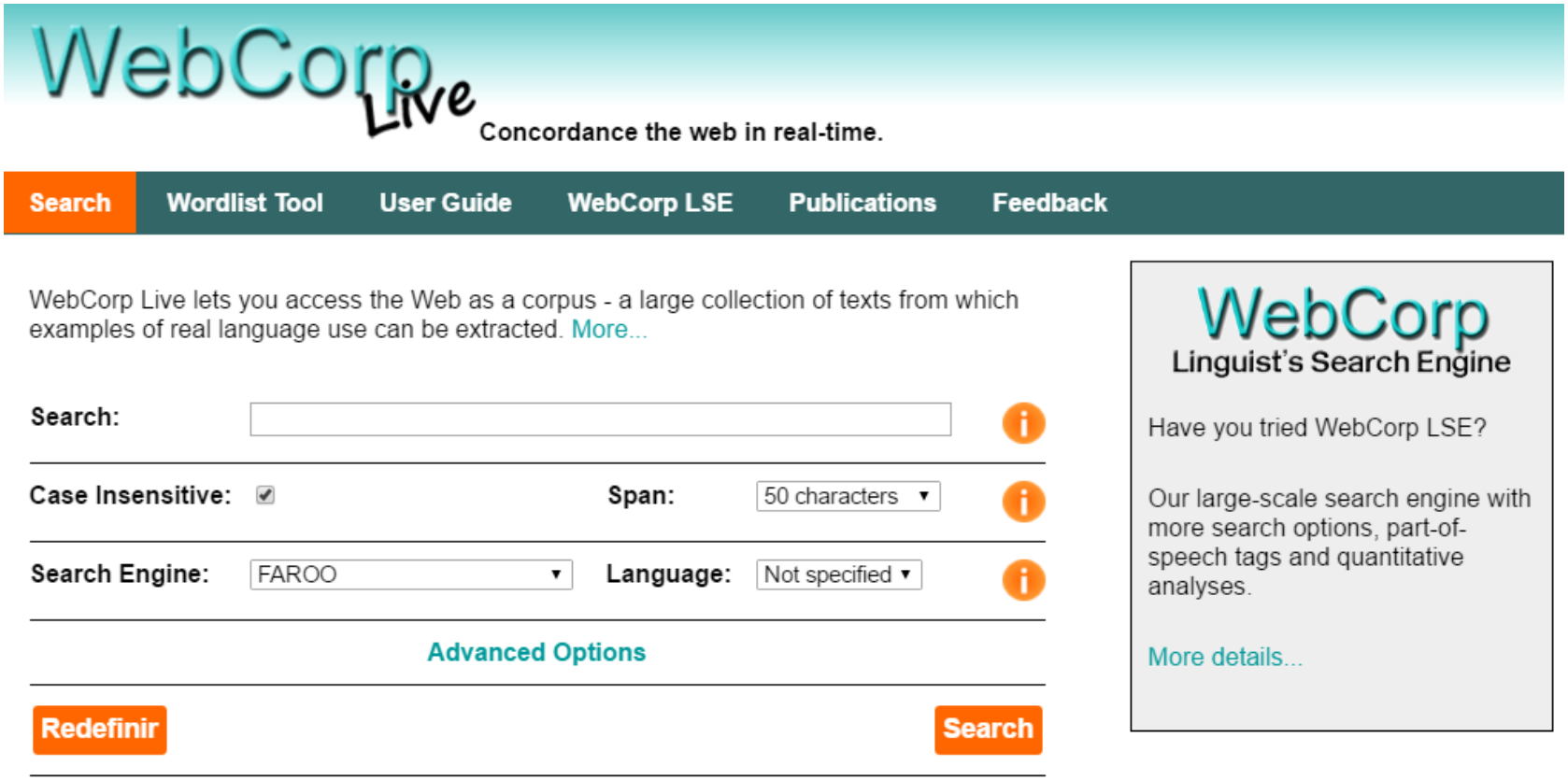
7 questões principais (Hovy e Lavid, 2010)

1. **Seleção do córpus** (e as questões já estudadas anteriormente)
   - Copyright
   - Disponibilidade
   - Gênero textual
   - Domínio
   - Tempo
   - Representatividade
   - Balanceamento
   - Web como córpus
   - Etc.

Como garantir representatividade e balanceamento?

WebCorp: The Web as Corpus

www.webcorp.org.uk/live/



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

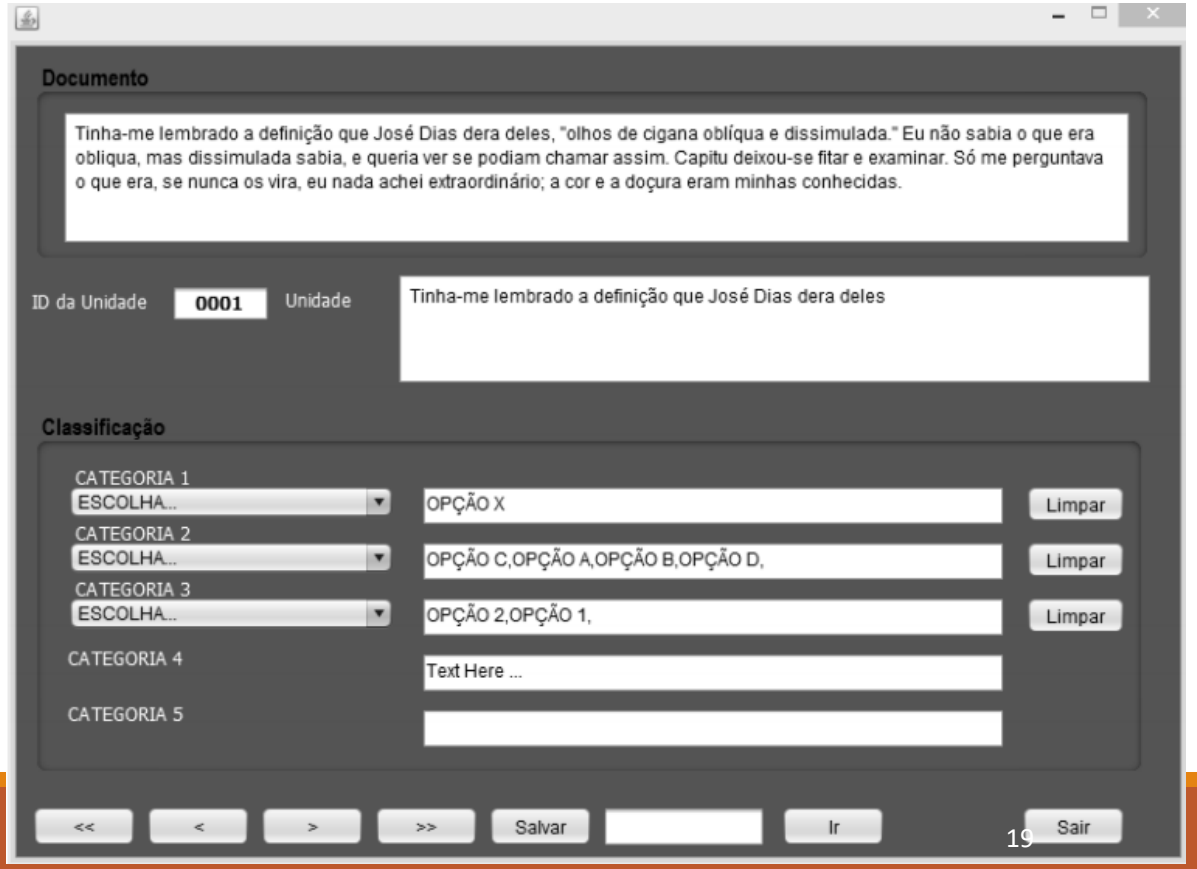
Ferramentas pré-fabricadas

Ferramentas customizáveis

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

Questões

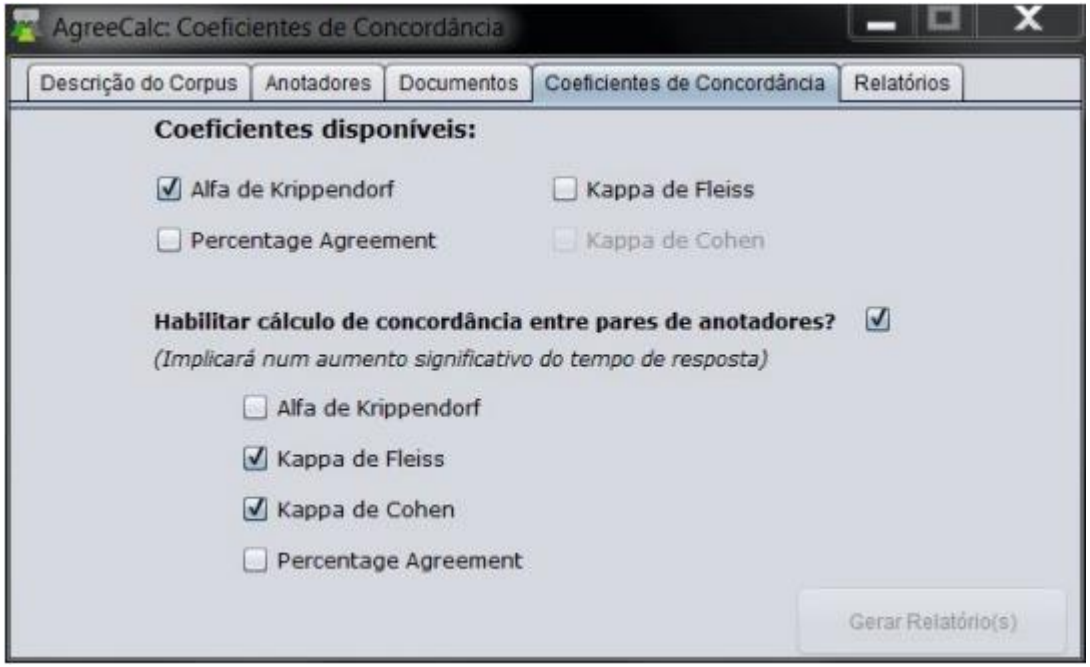
- Custo de customi-zação?
- A mais apropriada para a tarefa?



Ferramentas pré-fabricadas

Ferramentas customizáveis

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

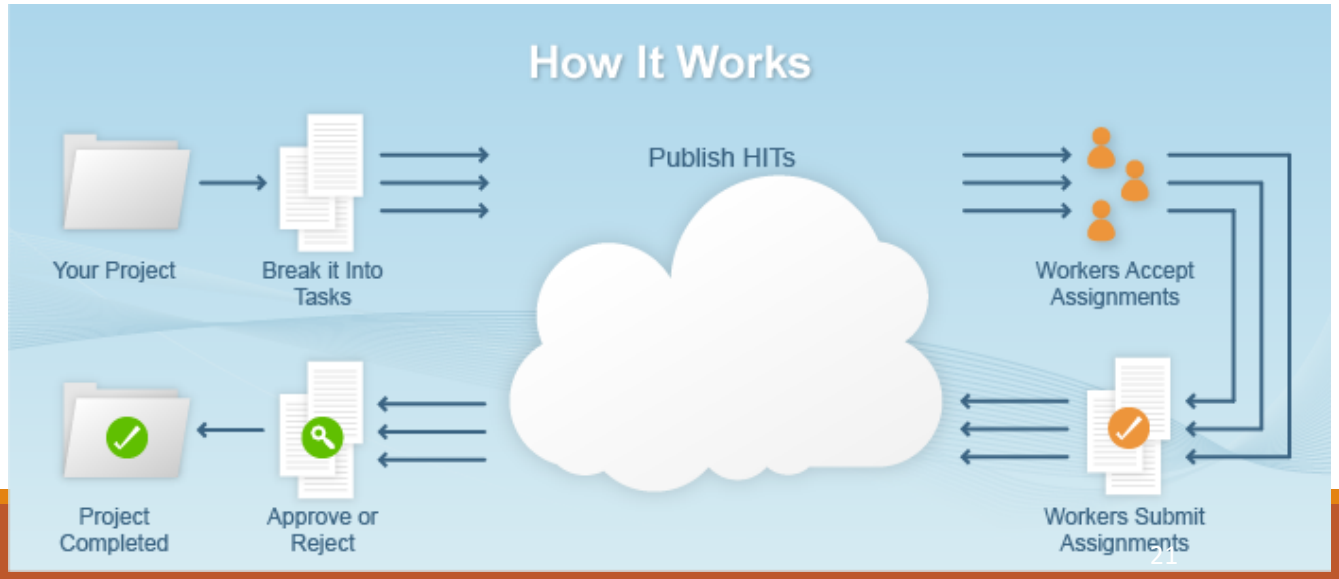


Crowdsourcing

Amazon Mechanical Turk

Questões (Fort et al., 2011)

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



Colaboração

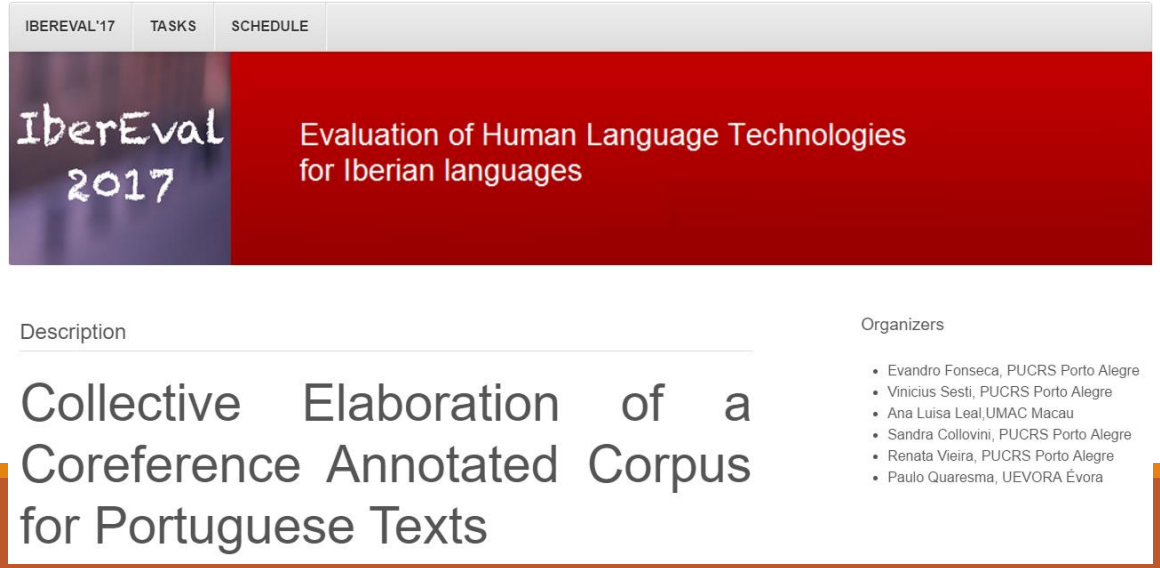
Anotação colaborativa, distribuída (escopo normalmente menor do que *crowdsourcing*, com mais “controle”)

Questões

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

Vantagens

- Dados
- Ferramentas



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

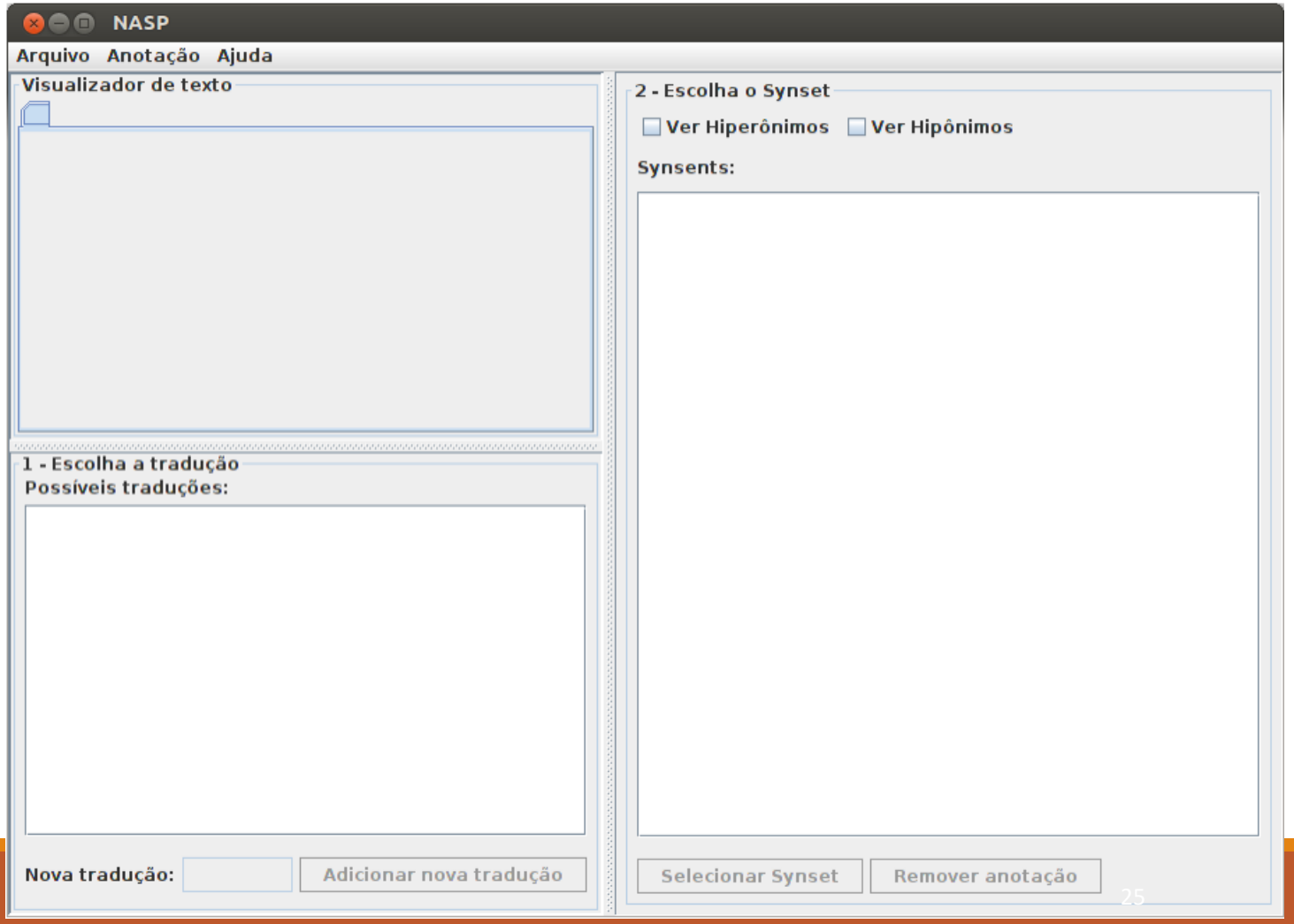
Criação e anotação

Exemplos reais

Sentido de palavras

NASP

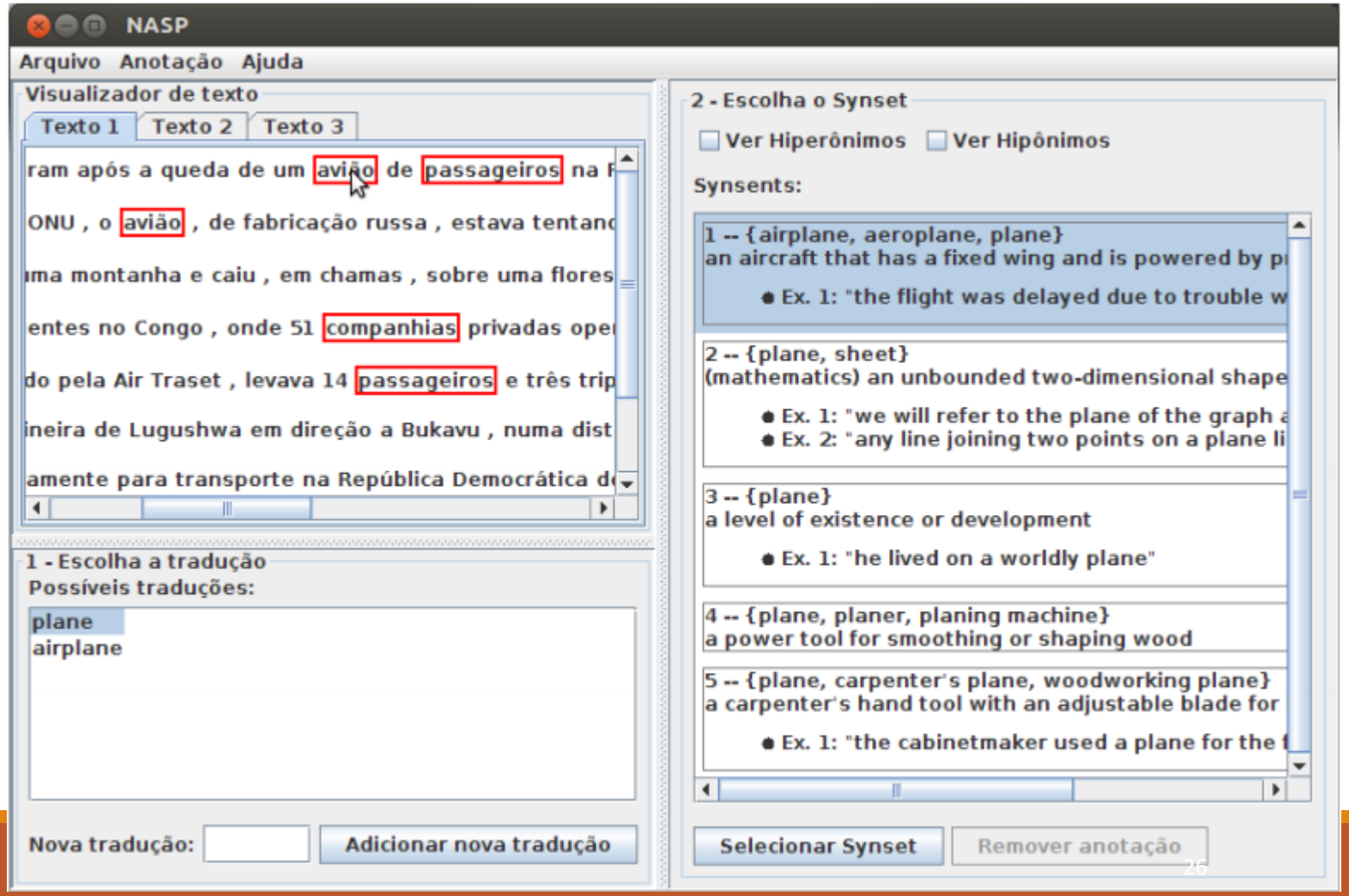
Nóbrega (2013)



Sentido de palavras

NASP

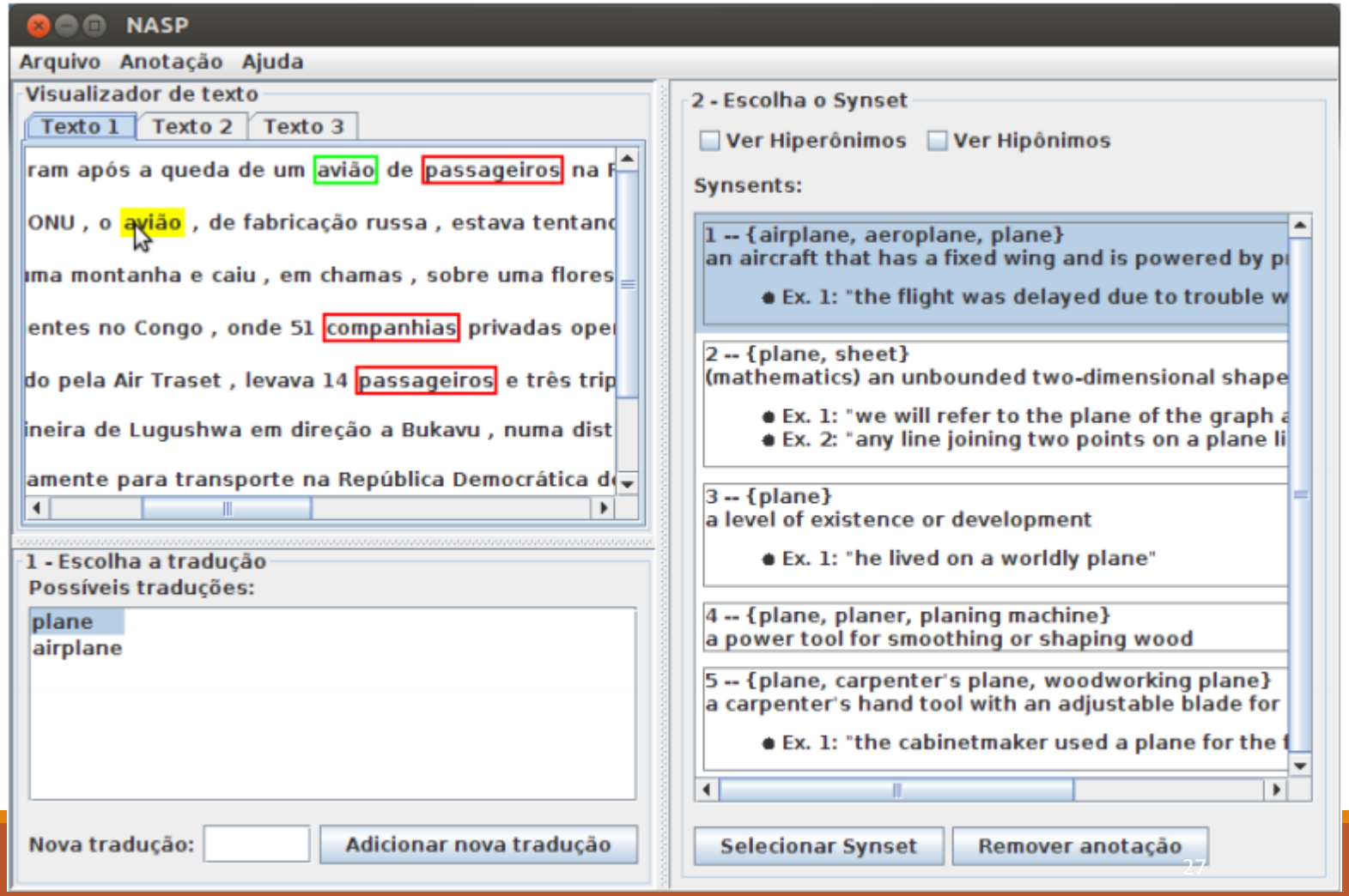
Nóbrega (2013)



Sentido de palavras

NASP

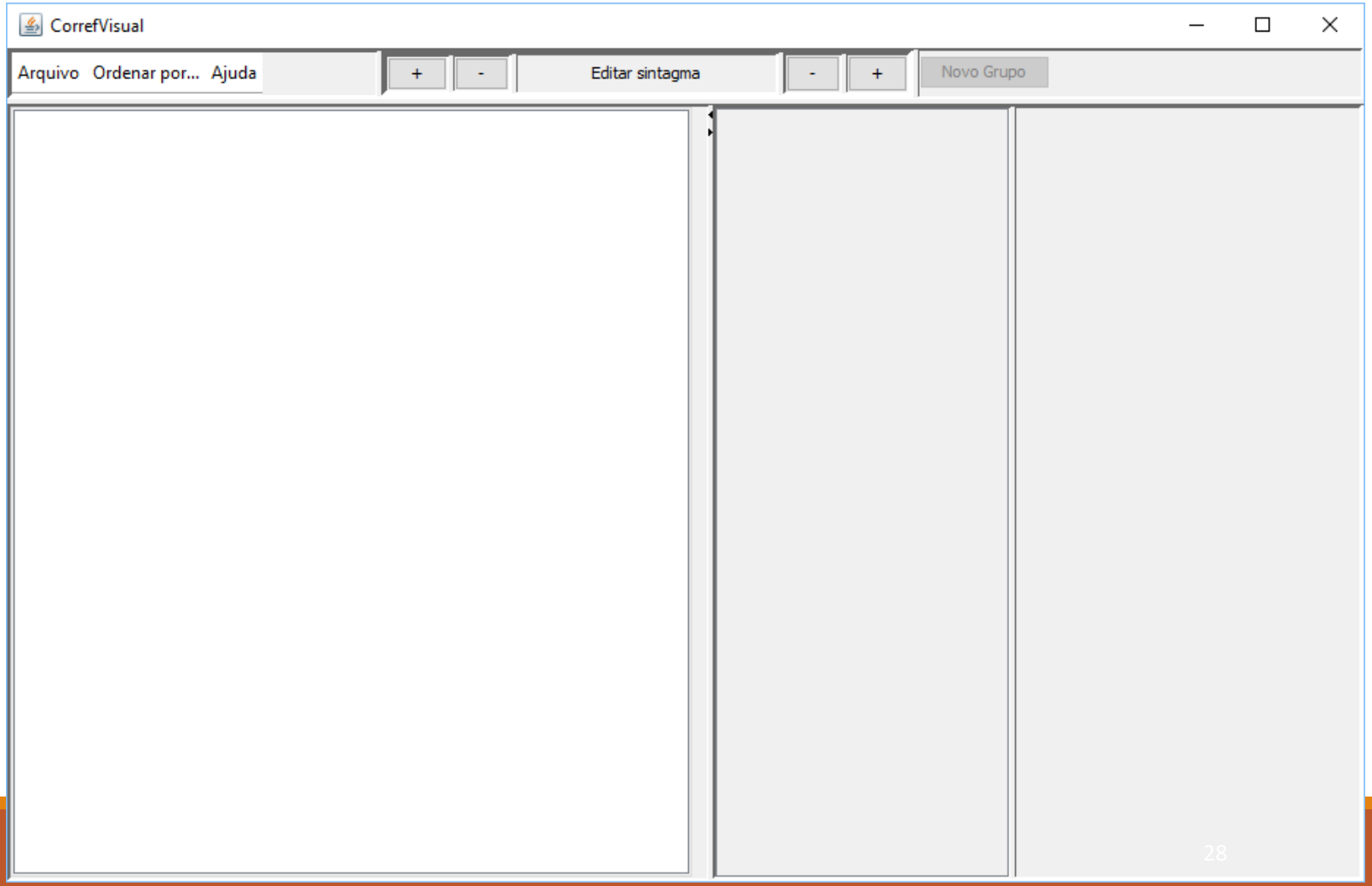
Nóbrega (2013)



Cadeias de correferência

CorrefVisual

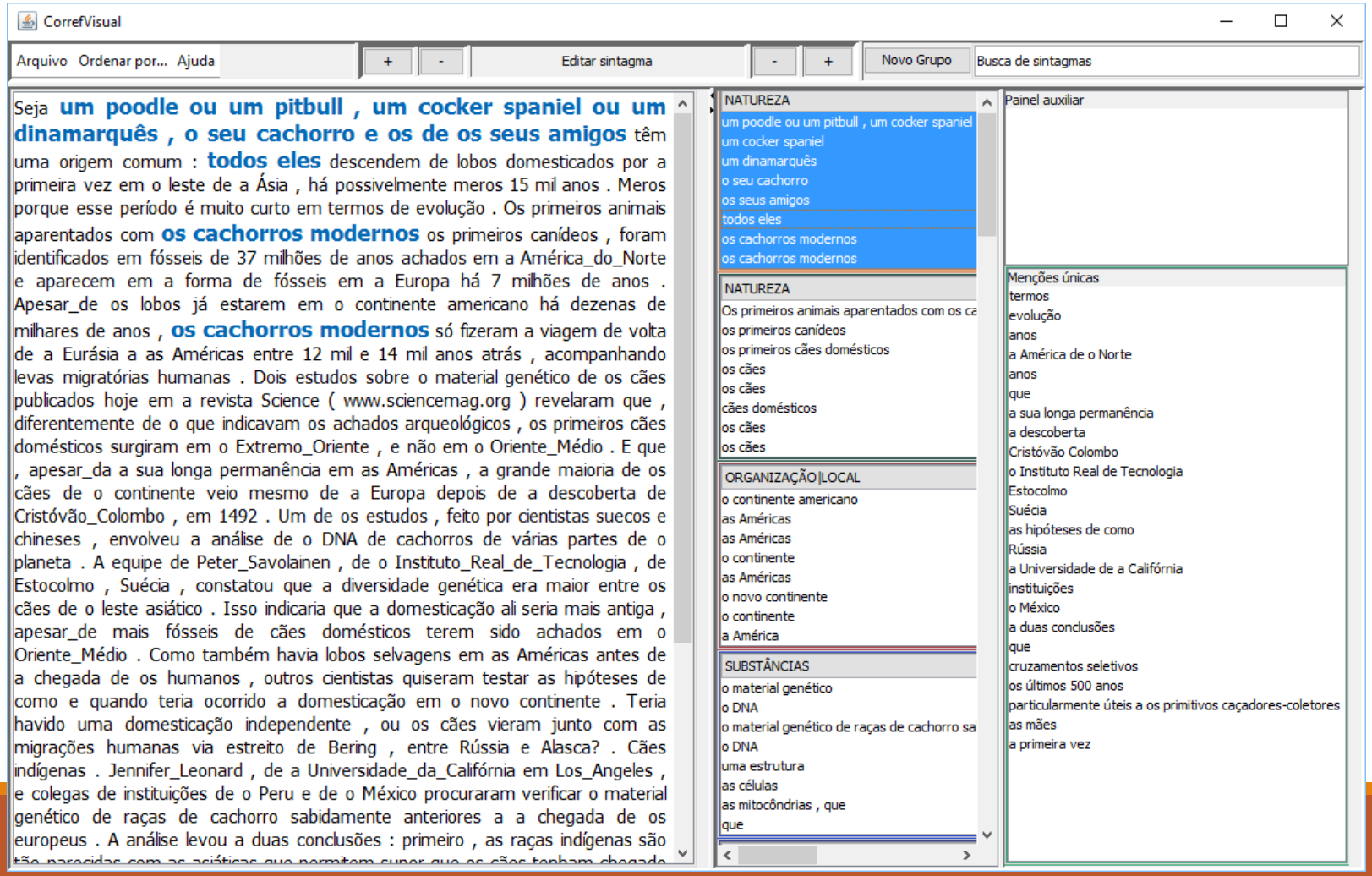
Fonseca et al. (2017)



Cadeias de correferência

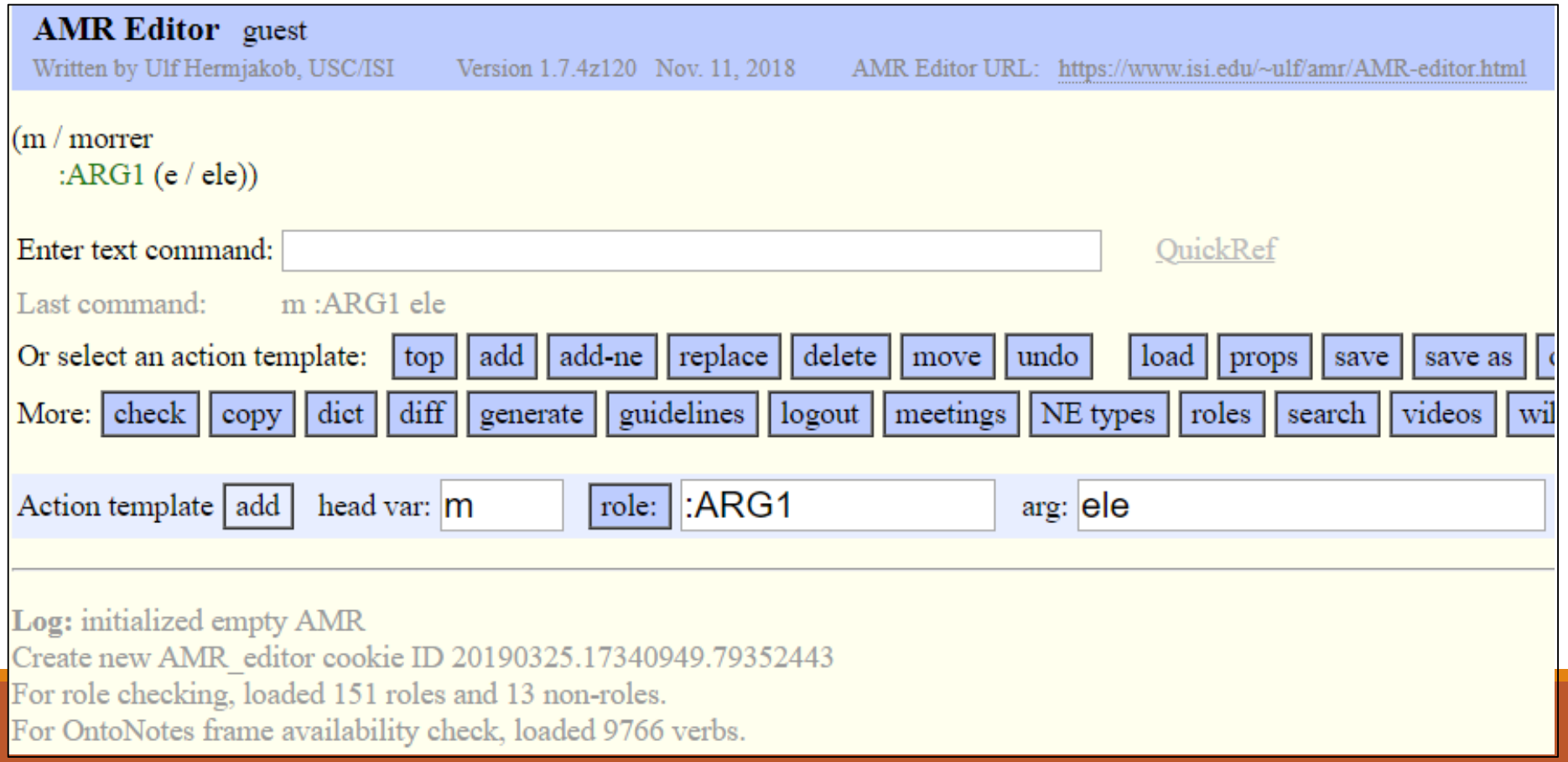
CorrefVisual

Fonseca et al. (2017)



Análise semântica

Abstract Meaning Representation (Banarescu et al., 2013)

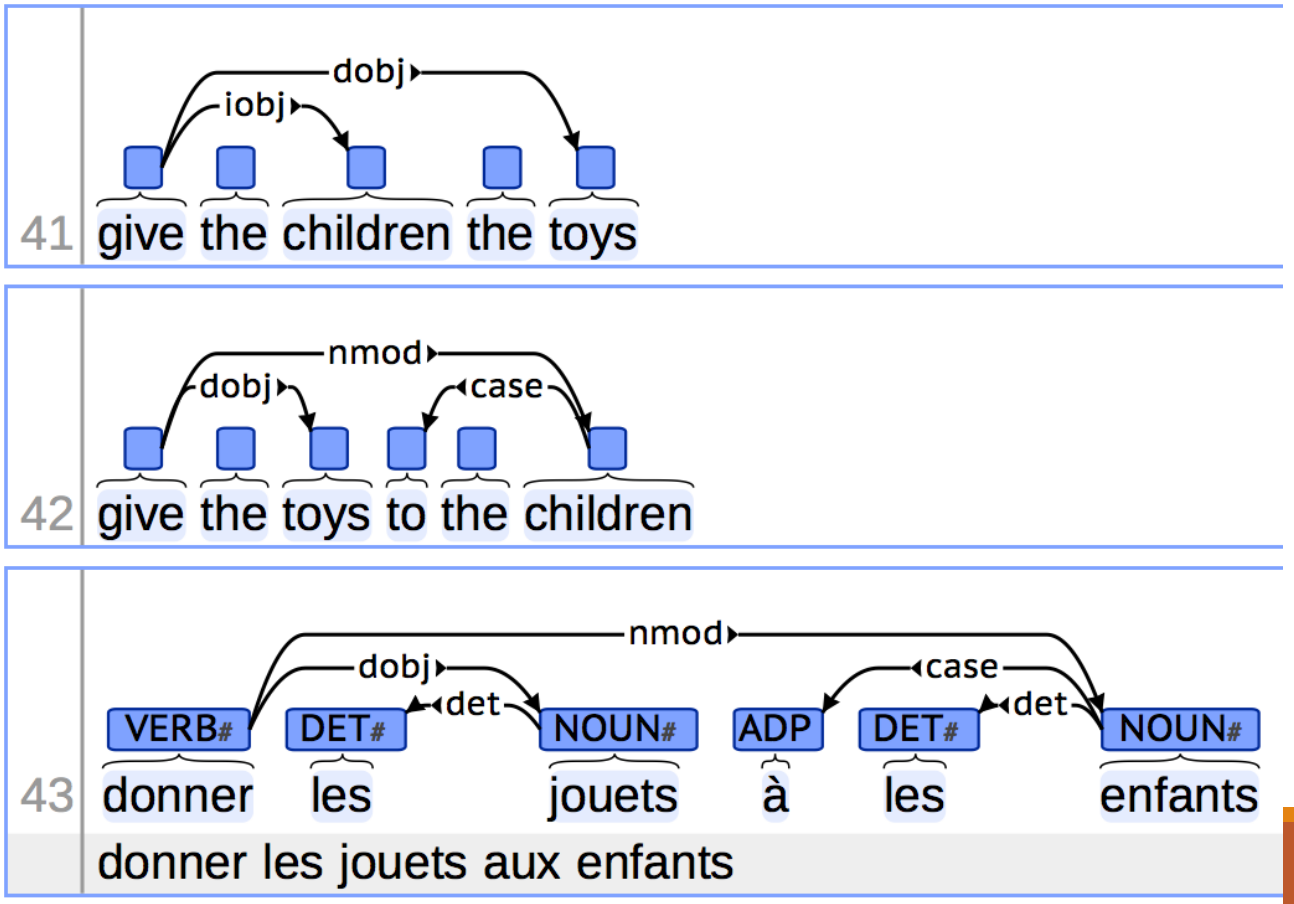


Polaridade de *tweets*

Brum (2017)

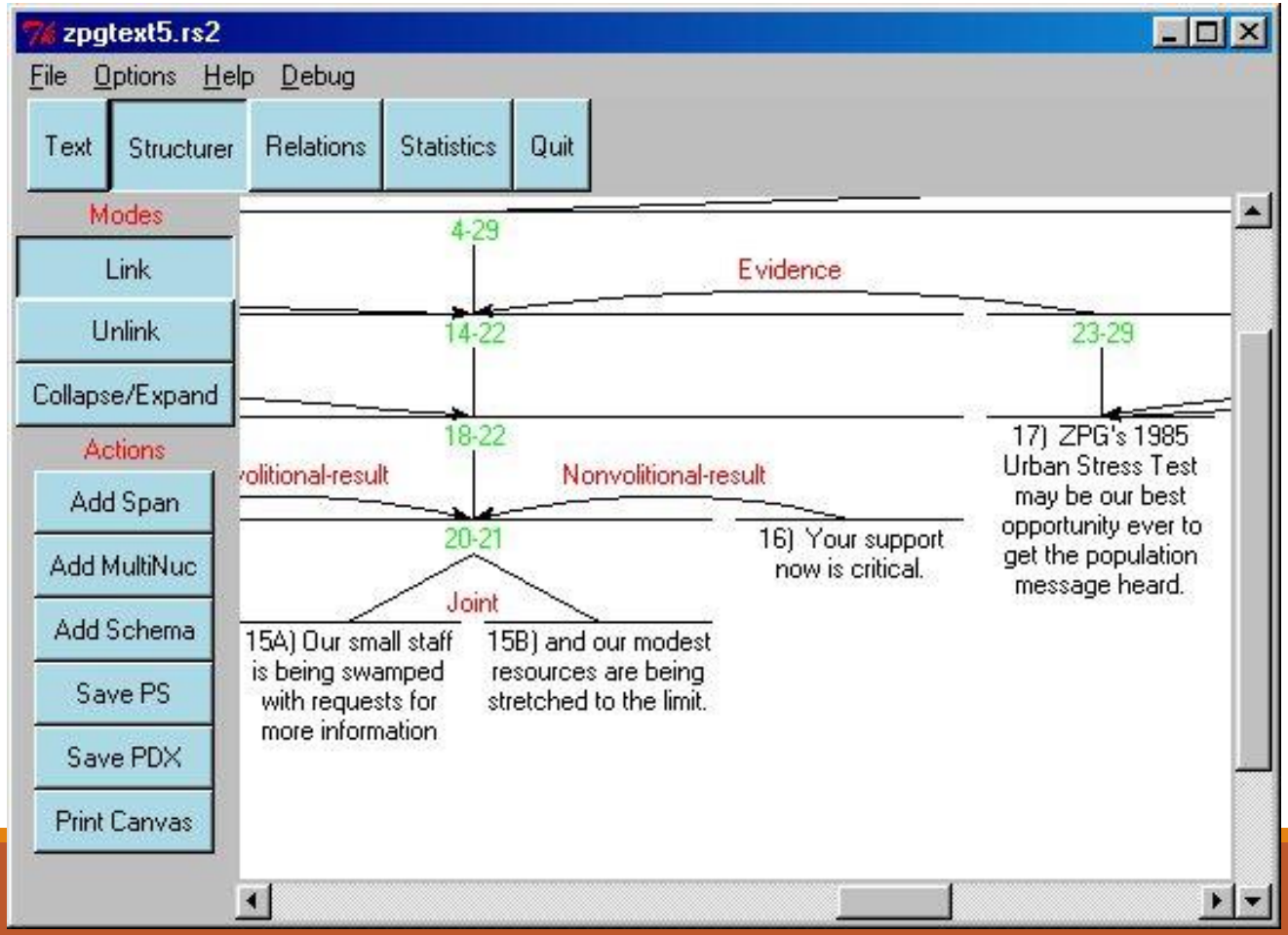
Dependências sintáticas

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



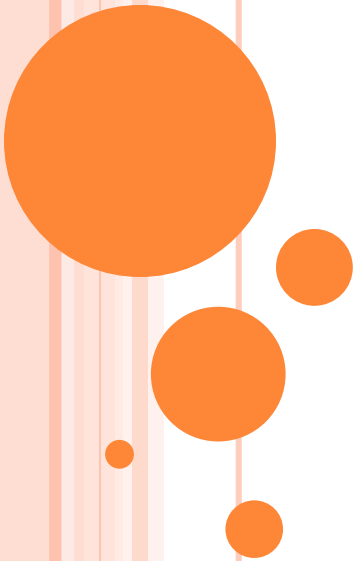
Relações retóricas

RSTTool (O'Donnell, 1997)



Caso real

O PROJETO SUCINTO



sucinto info

summarization
for clever information access

NILC ICMC USP SÃO CARLOS SCC

FAPESP

sucinto info

summarization
for clever information access

- www.icmc.usp.br/~taspardo/sucinto
  - Projeto longo: de 2009 a 2017
    - Financiamento FAPESP, principalmente
  - Foco em ***sumarização multidocumento***

sucinto info

summarization
for clever information access

- 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.

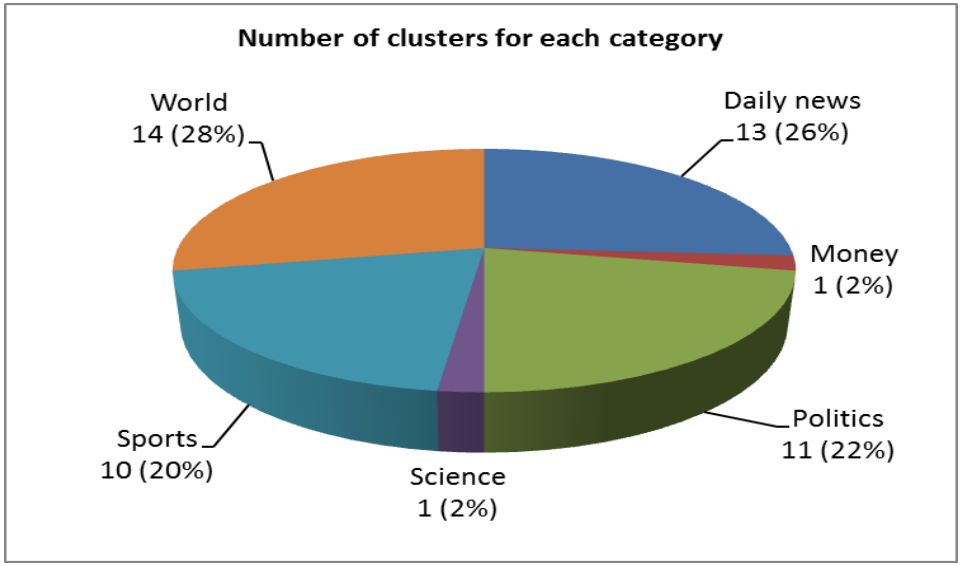
sucinto info

summarization
for clever information access

- www.icmc.usp.br/~taspardo/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 córpus 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 Córpus 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 Cabezudo, 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 Criacao de Novos Sumarios 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.

Agostini, V.; Camargo, R.T.; Pardo, T.A.S.; Di Felippo, A. (2012). Alinhamento manual de textos e sumários em um corpus jornalístico multidocumento. XI Encontro de Linguística de Corpus - ELC, pp. 1-5. September 13-15, 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

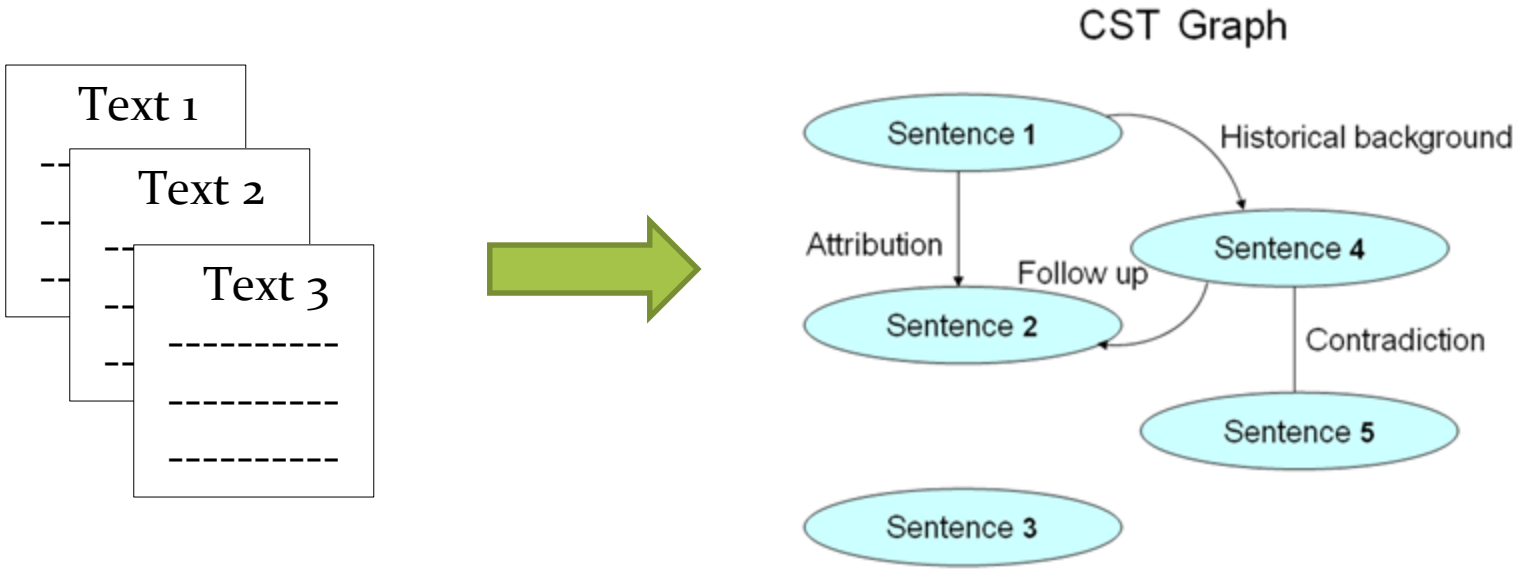
- Tipos e frequências de alinhamento

	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

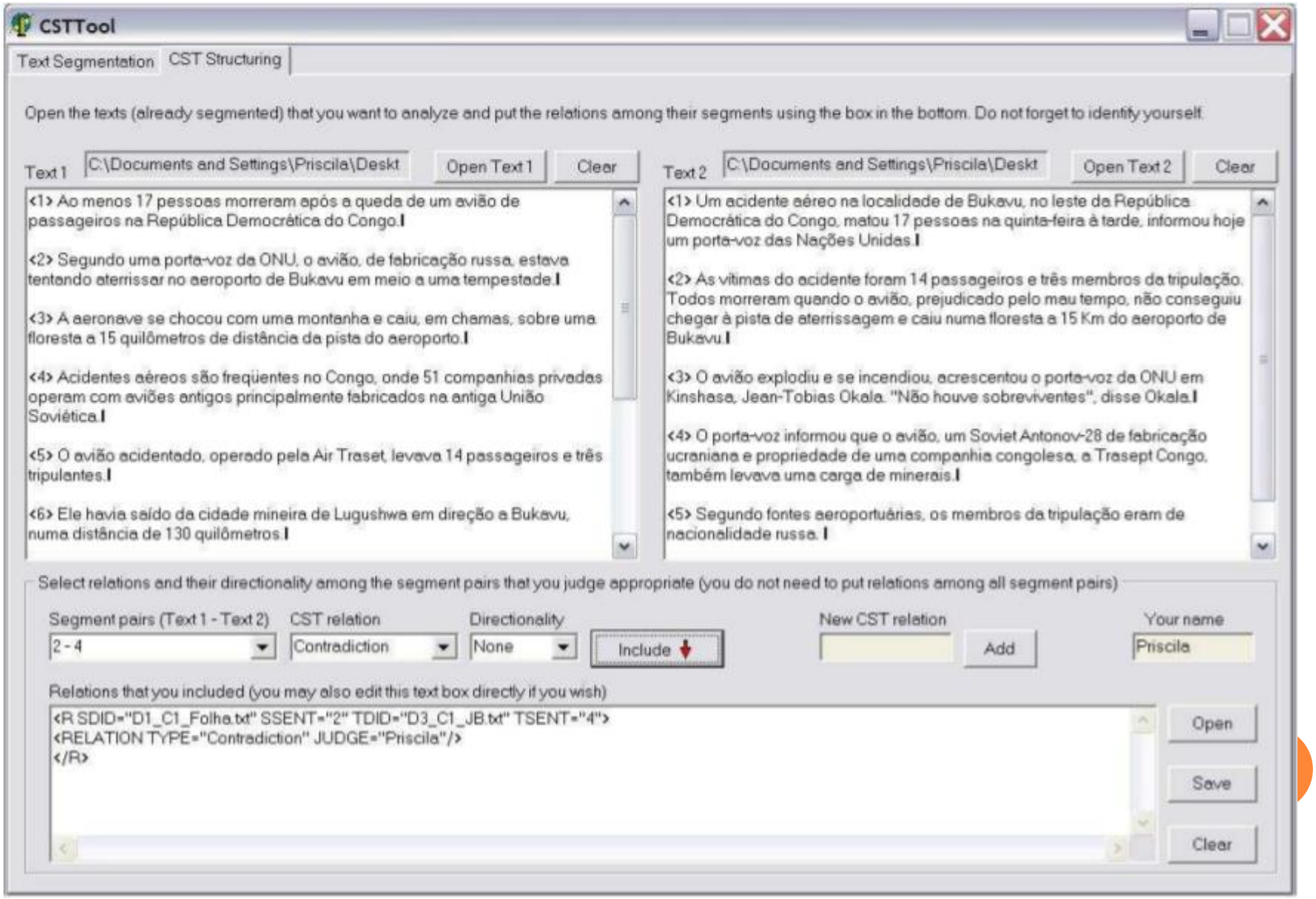


Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the Proceedings of the 3rd RST Brazilian Meeting, pp. 88-105. October 26, Cuiabá/MT, Brazil.

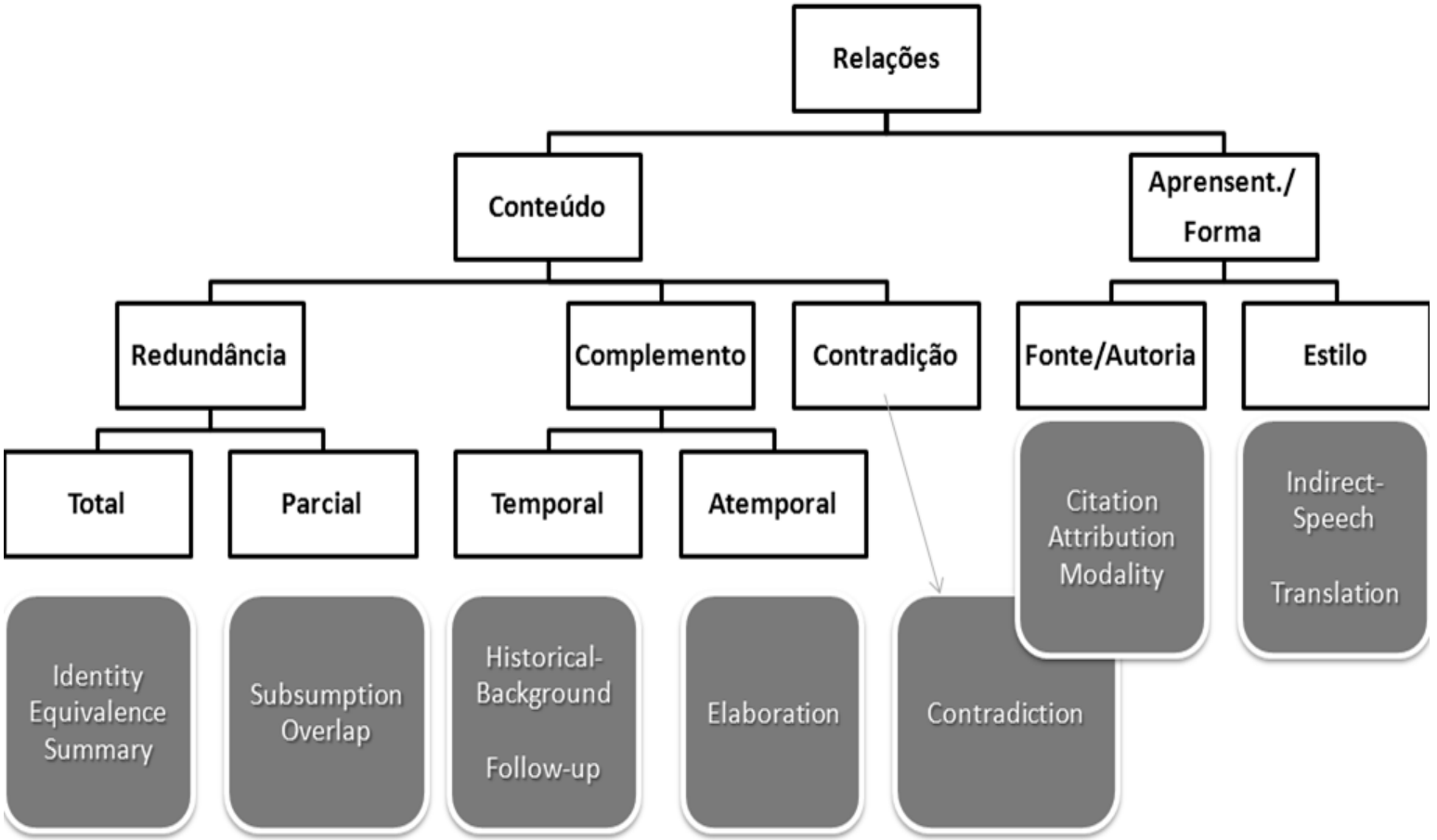
Cross-document Structure Theory (CST) (Radev, 2000)

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

CSTTool


CSTTool
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\Deskt
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 freqüentes 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\Deskt
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 congolesa, 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)
2 - 4
CST relation
Contradiction
Directionality
None
Include
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

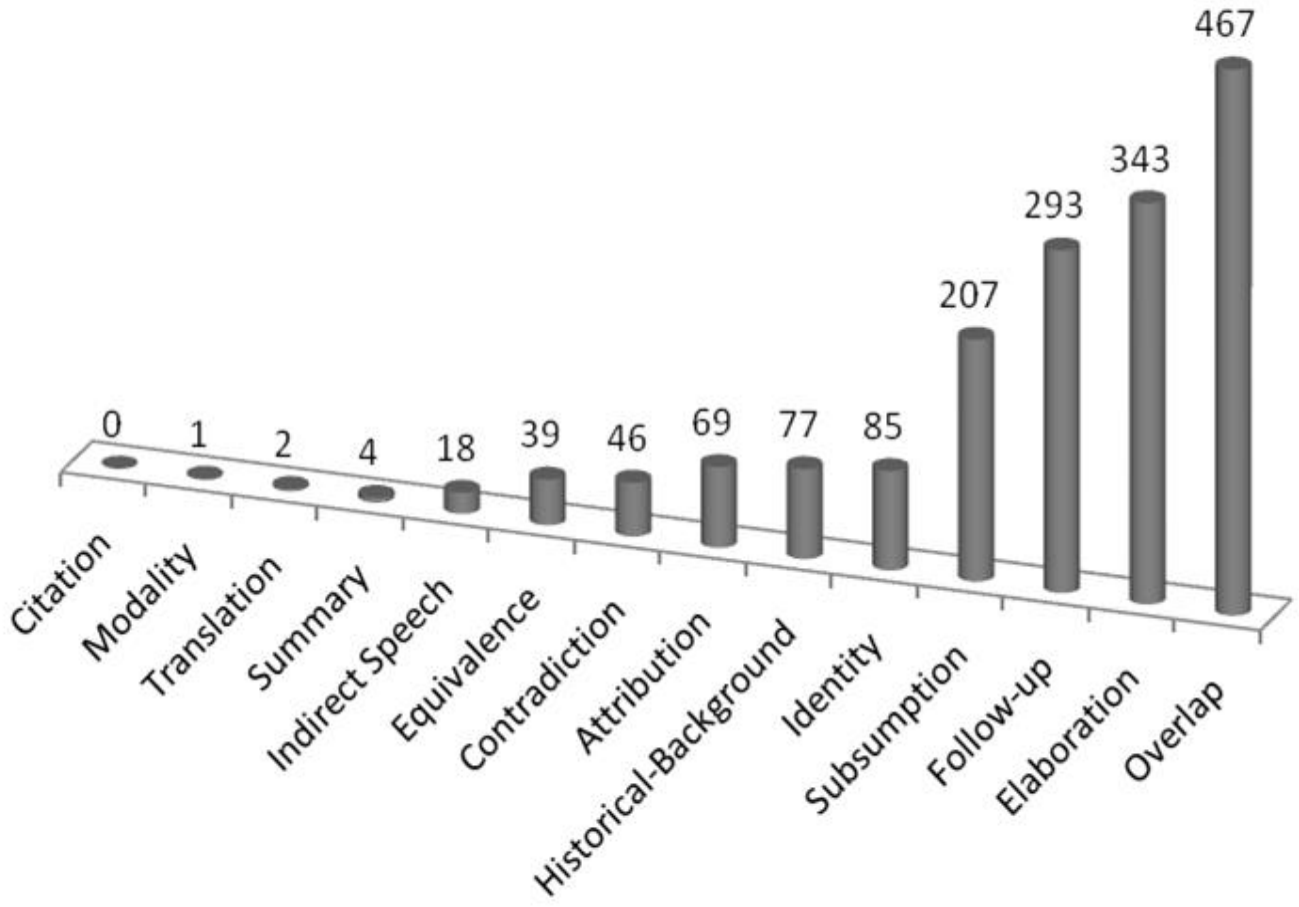
TIPOLOGIA DE RELAÇÕES (MAZIERO ET AL., 2014)



Maziero, E.G.; Castro Jorge, M.L.R.; Pardo, T.A.S. (2014). Revisiting Cross-document Structure Theory for multi-document discourse parsing. Information Processing & Management, Vol. 50, N. 2, pp. 297-314.

CST

- Fenômenos multidocumento



CST

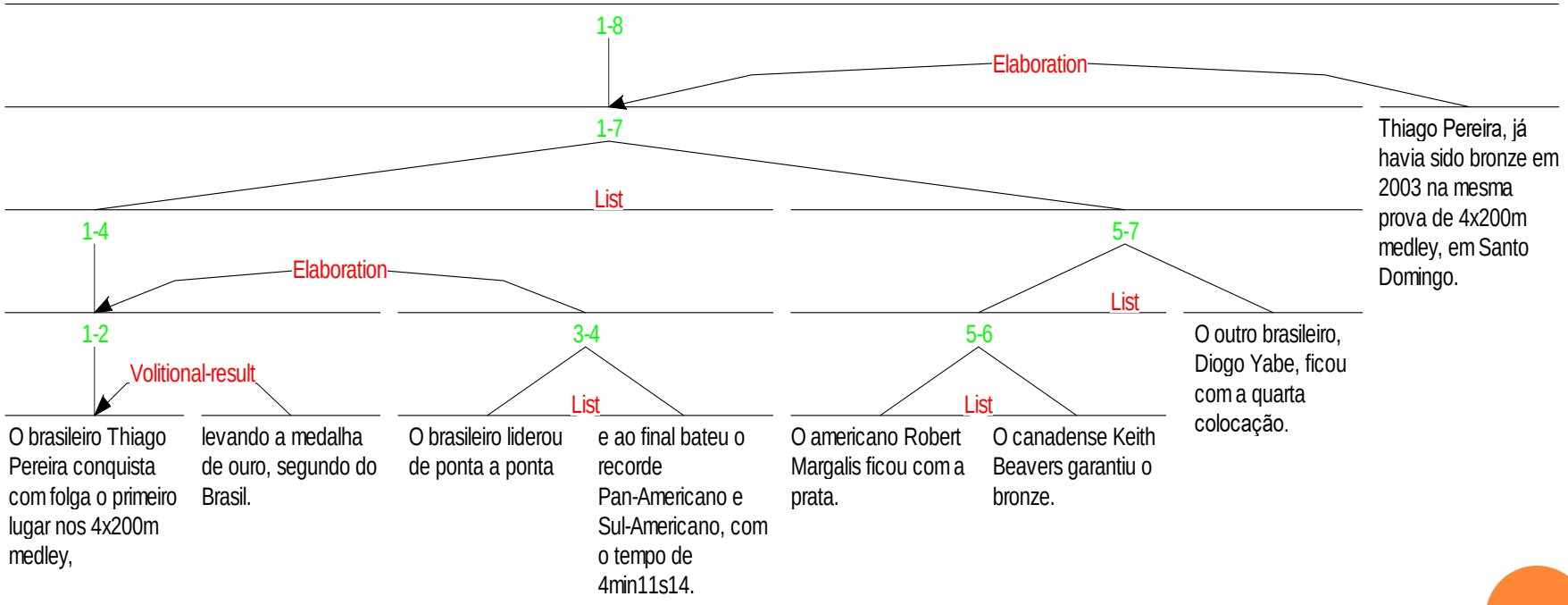
- Confiança na anotação

	Kappa	*Porcentagem de concordância*		
		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 e Thompson, 1987)

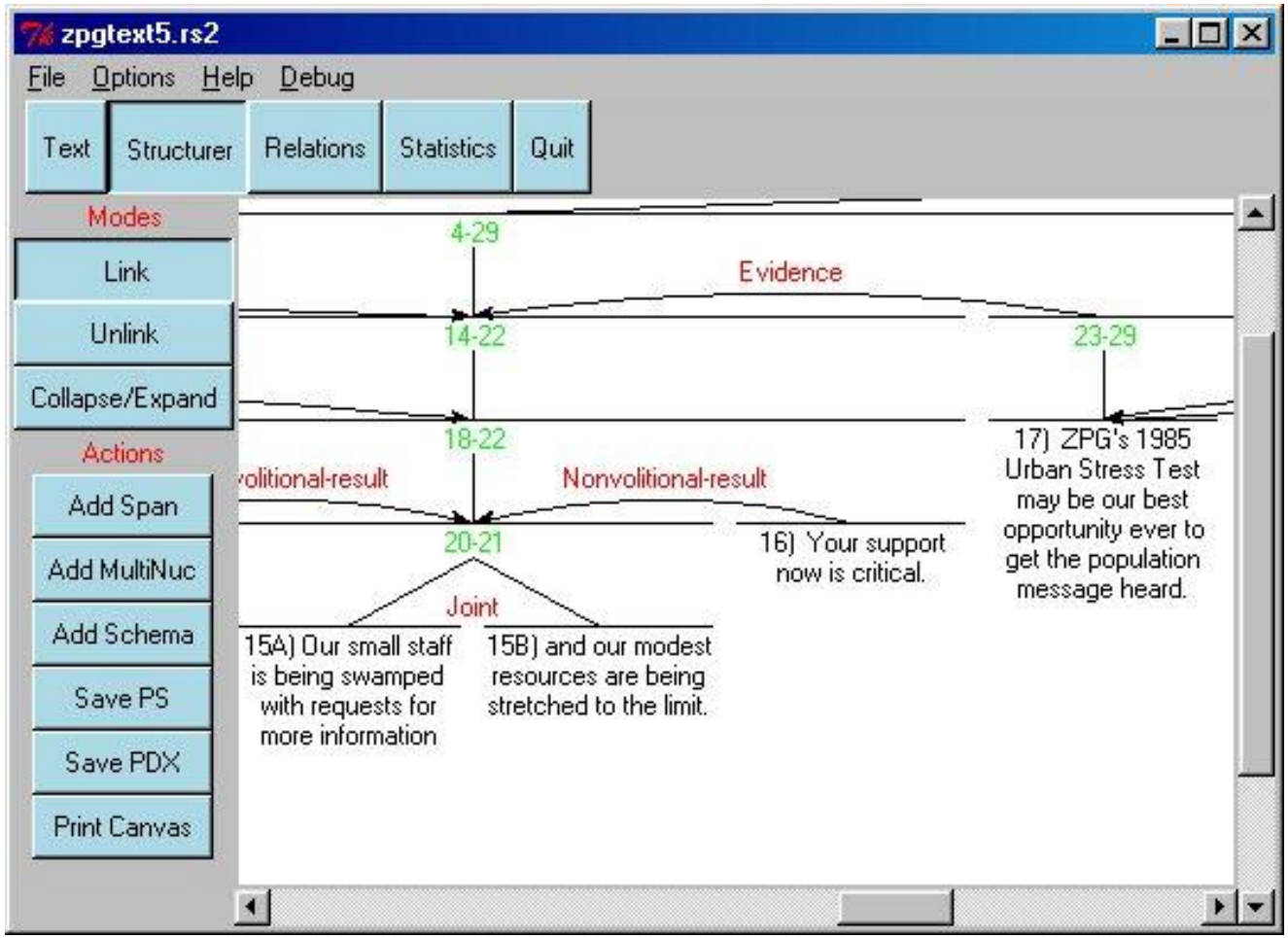
- Anotação de córpus
  - 8 pessoas
  - Quase 1 semestre de anotação
  - RSTTool

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

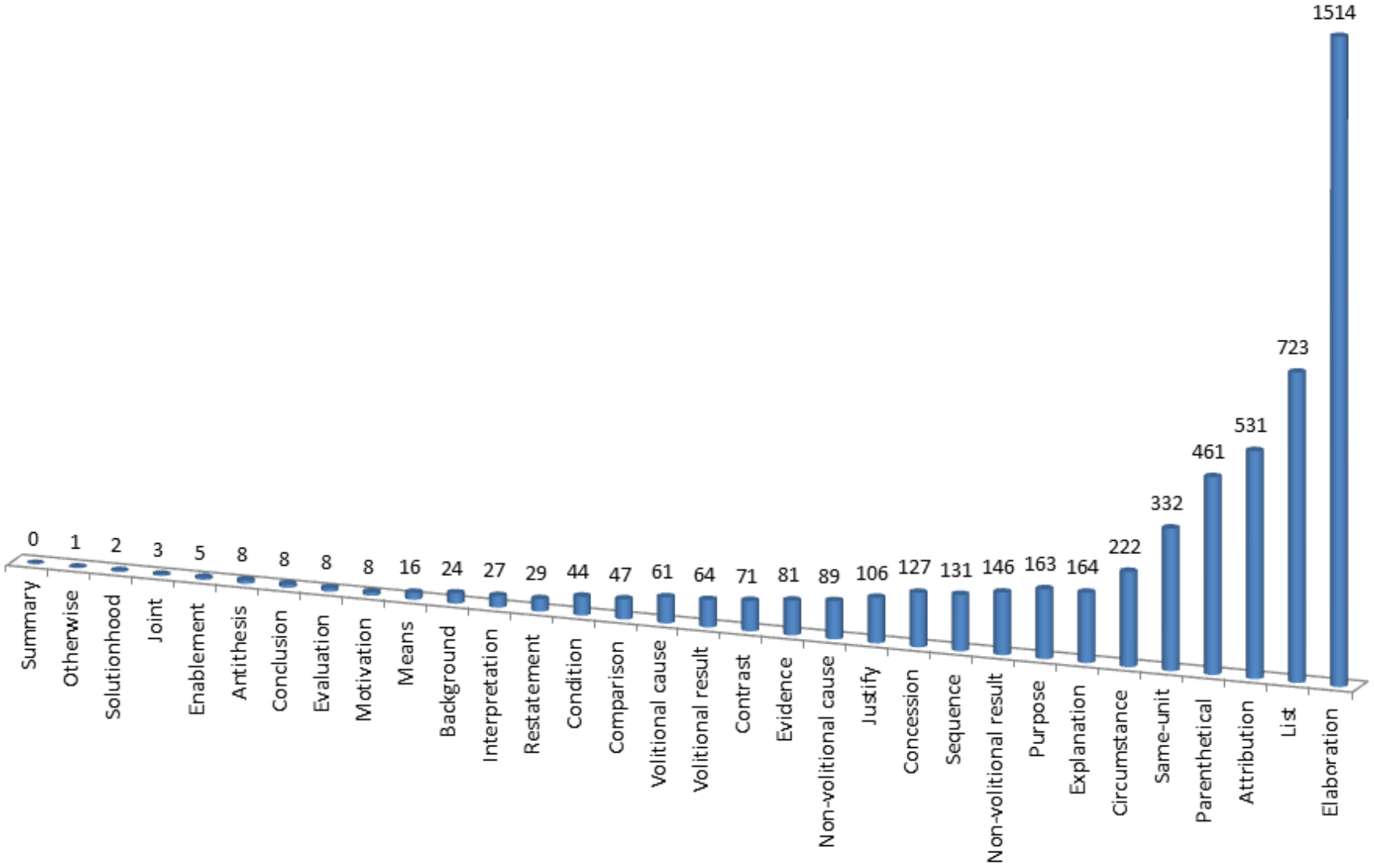
Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the Proceedings of the 3rd RST Brazilian Meeting, pp. 88-105. October 26, Cuiabá/MT, Brazil.

RSTTool

- Editor clássico (O'Donnell, 2000)



RST


Summary 0
Otherwise 1
Solutionhood 2
Joint 3
Enablement 5
Antithesis 8
Conclusion 8
Evaluation 8
Motivation 8
Means 16
Background 24
Interpretation 27
Restatement 29
Condition 44
Comparison 47
Volitional cause 61
Volitional result 64
Contrast 71
Evidence 81
Non-volitional cause 89
Justify 106
Concession 127
Sequence 131
Non-volitional result 146
Purpose 163
Explanation 164
Circumstance 222
Same-unit 332
Parenthetical 461
Attribution 531
List 723
Elaboration 1514

RST

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

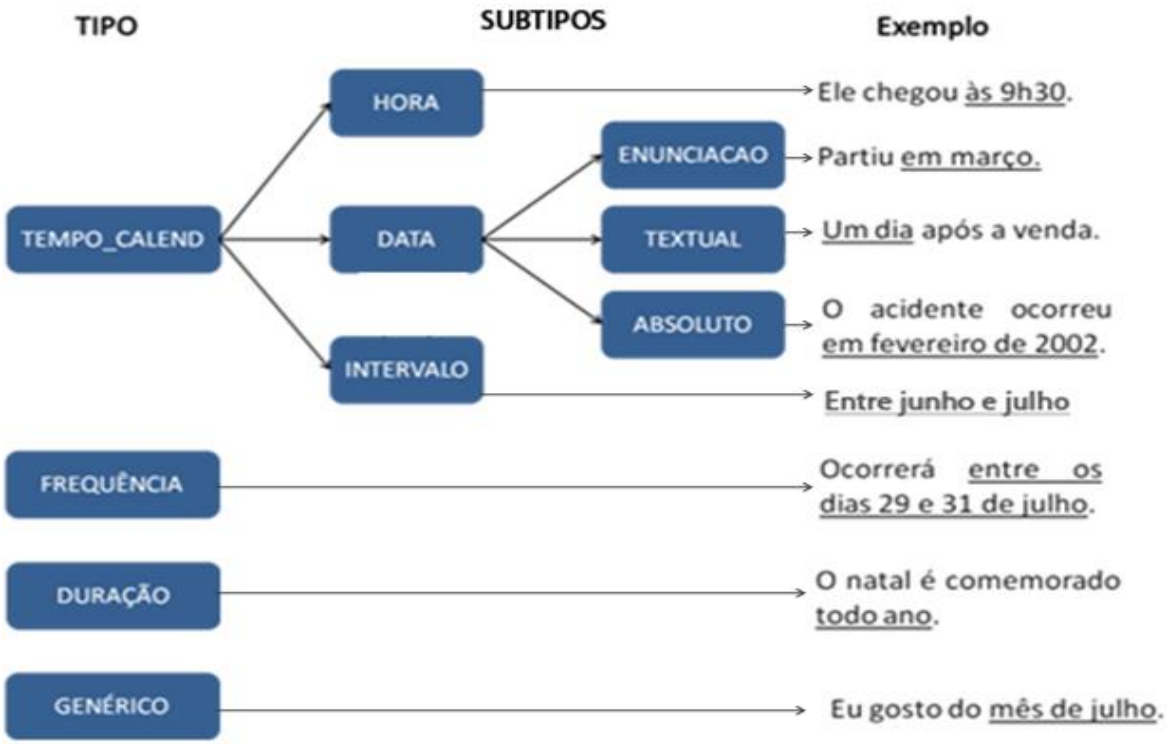
Evaluated Criteria	***Precision***	***Recall***	***F-Measure***
Simple Textual Segments	*0.91*	*0.91*	*0.91*
Complex Textual Segments	*0.78*	*0.78*	*0.78*
Nuclearity	*0.78*	*0.78*	*0.78*
Relations	*0.66*	*0.66*	*0.66*

Maziero, E.G. and Pardo, T.A.S. (2009). Automatização de um Método de Avaliação de Estruturas Retóricas. In the Proceedings of the RST Brazilian Meeting, pp. 1-9. September 10, São Carlos/SP, Brazil

Menezes Filho, L.A. and Pardo, T.A.S. (2011). Detecção de Expressões Temporais no Contexto de Sumarização Automática. In the Proceedings of the 2nd STIL Student Workshop on Information and Human Language Technology, pp. 1-3. October 24-25, Cuiabá/MT, Brazil.

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 (...)."*

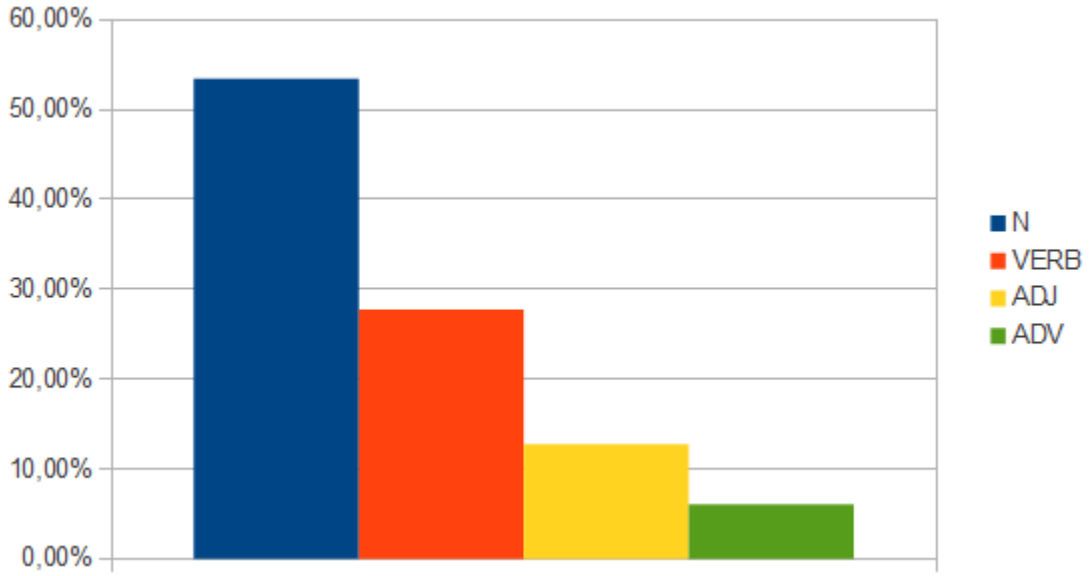
Frequência no córpus

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

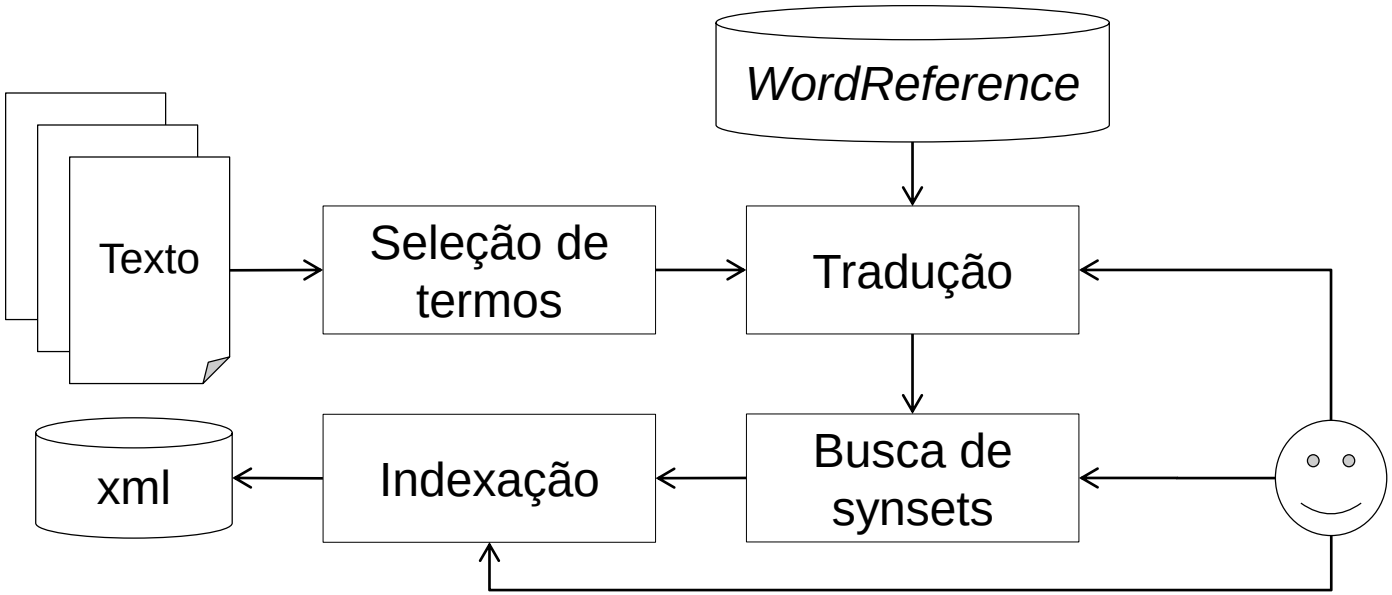


Nóbrega, F.A.A. and Pardo, T.A.S. (2014). General Purpose Word Sense Disambiguation Methods for Nouns in Portuguese. In the Proceedings of the PROPOR 2014 PhD and MSc/MA Dissertation Contest / 11st International Conference on Computational Processing of Portuguese - PROPOR (LNAI 8775), pp. 94-101. October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

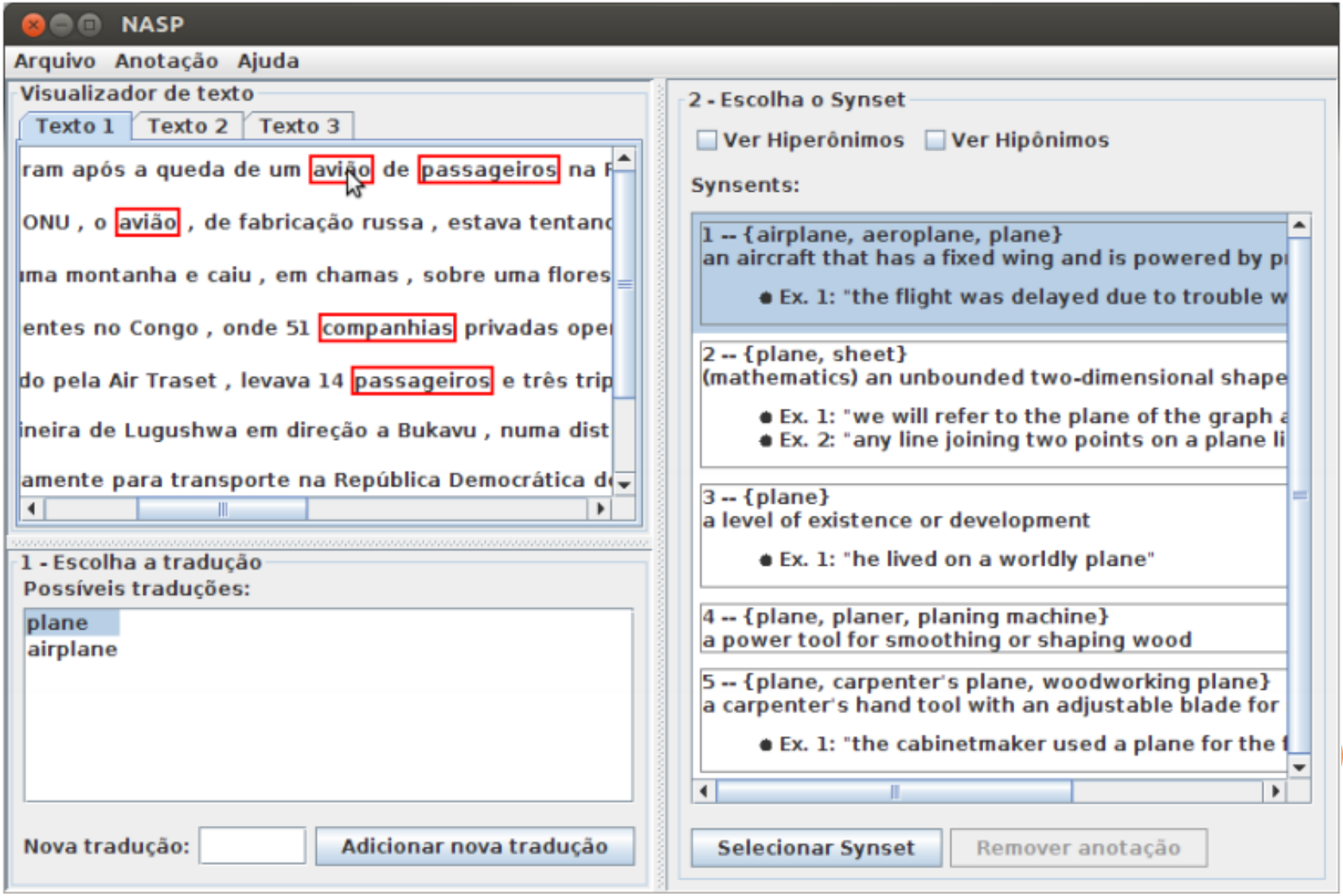
Sobrevilla Cabezudo, M.A.; Maziero, E.G.; Souza, J.W.C.; Dias, M.S.; Cardoso, P.C.F.; Balage Filho, P.P.; Agostini, V.; Nóbrega, F.A.A.; Barros, C.D.; Di Felippo, A.; Pardo, T.A.S. (2015). Anotação de Sentidos de Verbos em Textos Jornalísticos do Corpus CSTNews. Revista de Estudos da Linguagem - RELIN,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

Sentido

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



NASP


NASP
Arquivo Anotação Ajuda
Visualizador de texto
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
ıma montanha e caiu , em chamas , sobre uma flores
entes no Congo , onde 51 companhias privadas ope
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
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
Selecionar Synset
Remover anotação

Sentido

- Concordância

Kappa

	Substantivos	Verbos
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 ;-)

Cardoso, P.C.F.; Pardo, T.A.S.; Taboada, M. (2017). Subtopic annotation and automatic segmentation for news texts in Brazilian Portuguese. Corpora, Vol. 12, N. 1, pp. 23-54. Edinburgh University Press.

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

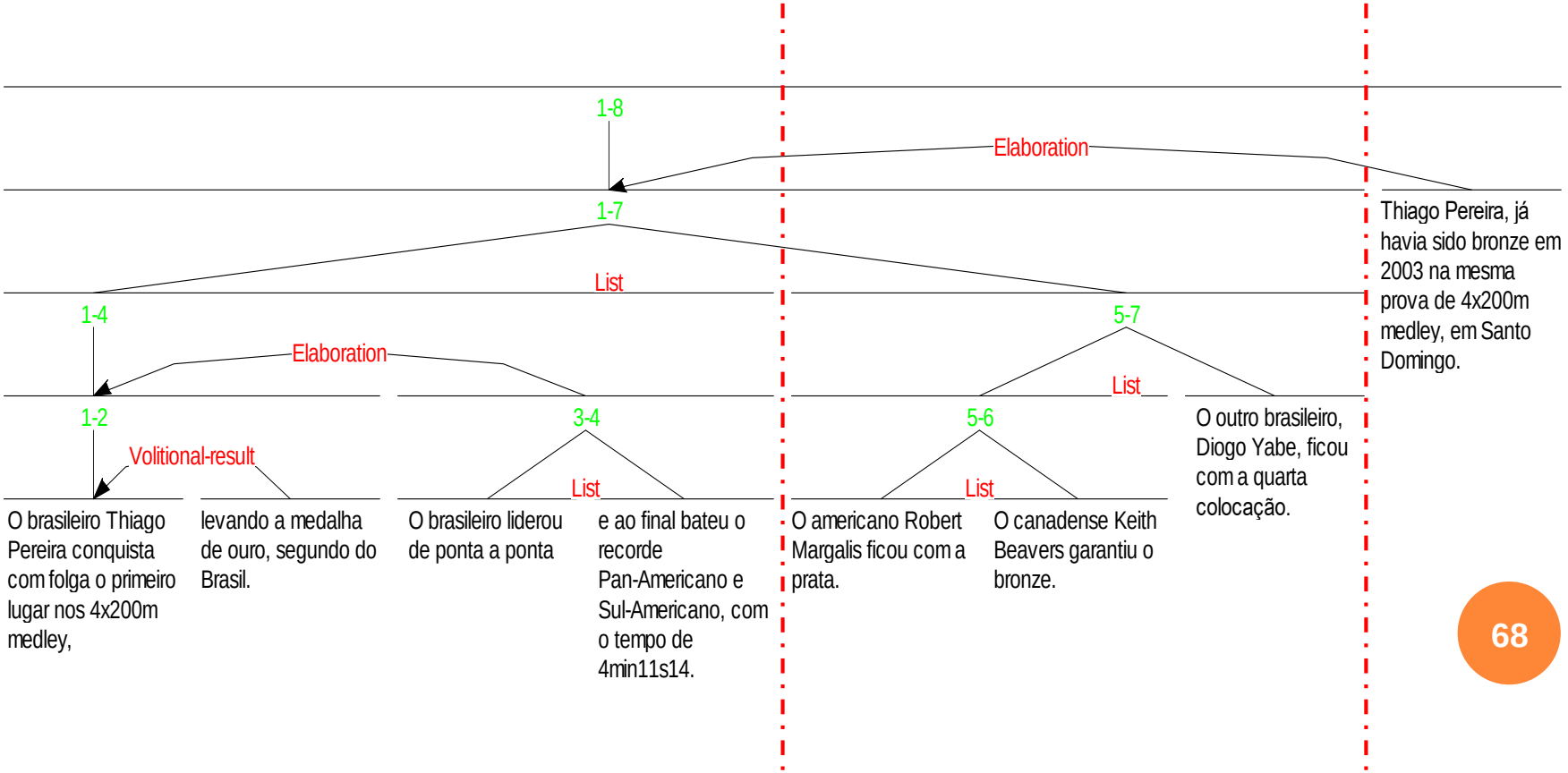
- Concordância

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**

Cardoso, P.C.F.; Pardo, T.A.S.; Taboada, M. (2013). On the contribution of discourse to topic segmentation. In the Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue – SIGdial, pp. 92-96. August 22-23, Metz/France.

RST & SUBTÓPICOS

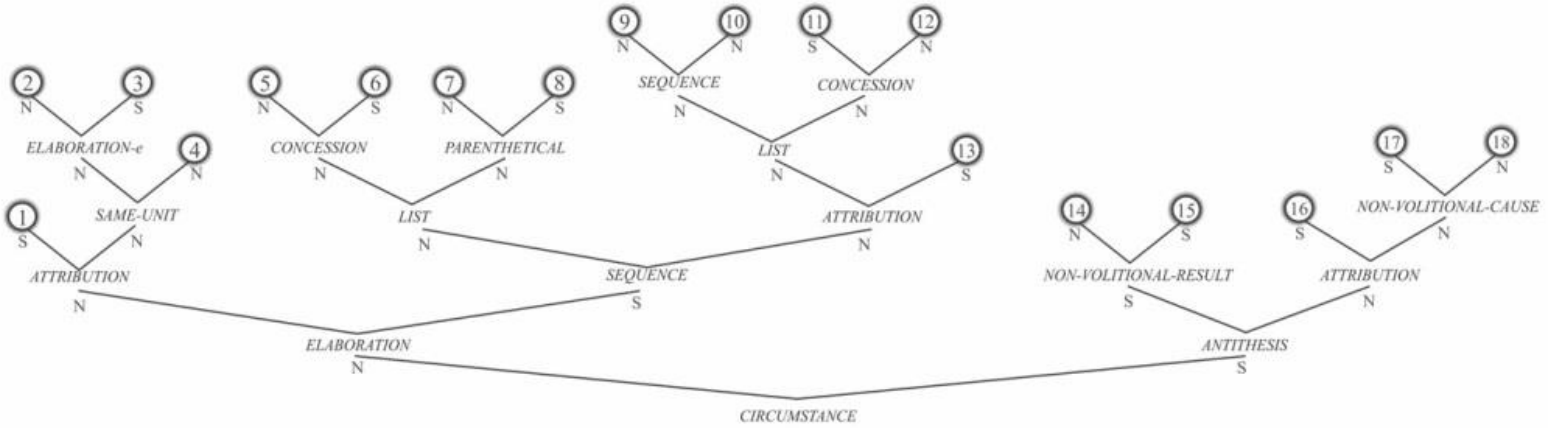
- Correspondência entre subtópicos e estruturação discursiva

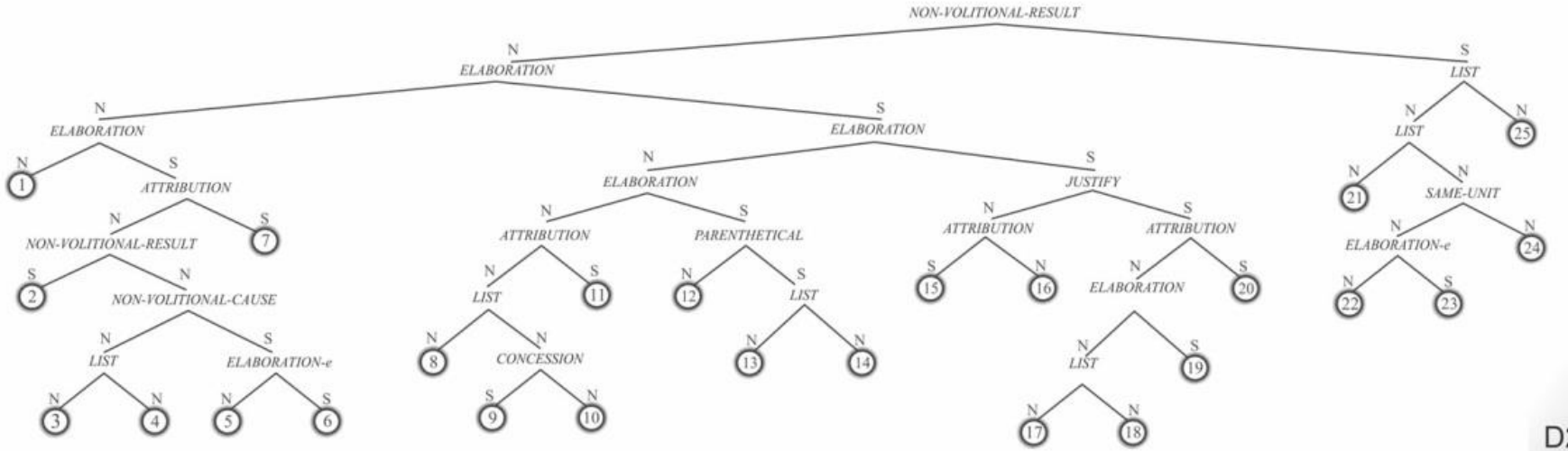


Por que parar nisso?

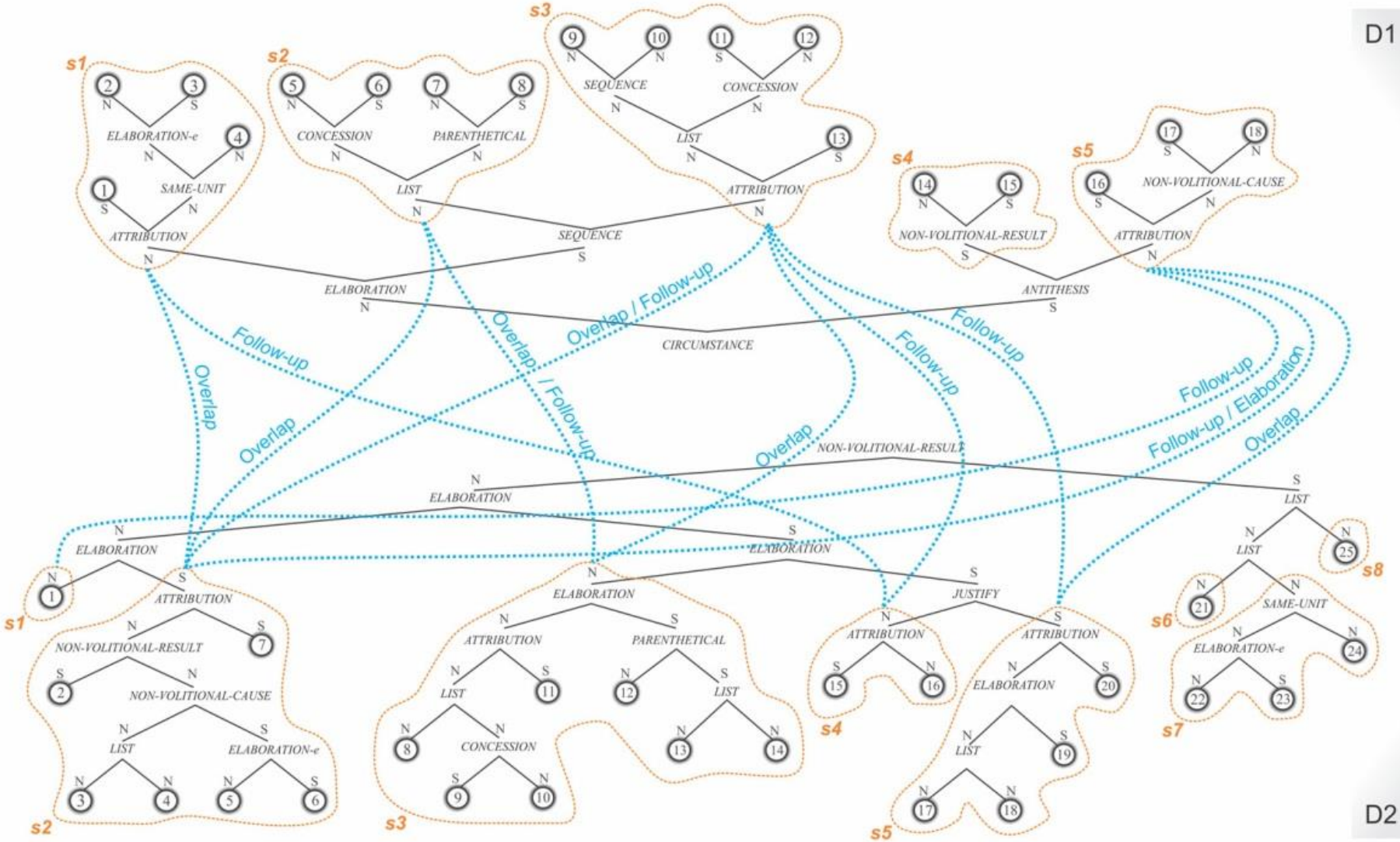
- O grande diferencial do CSTNews: correlação entre níveis até então inexplorada

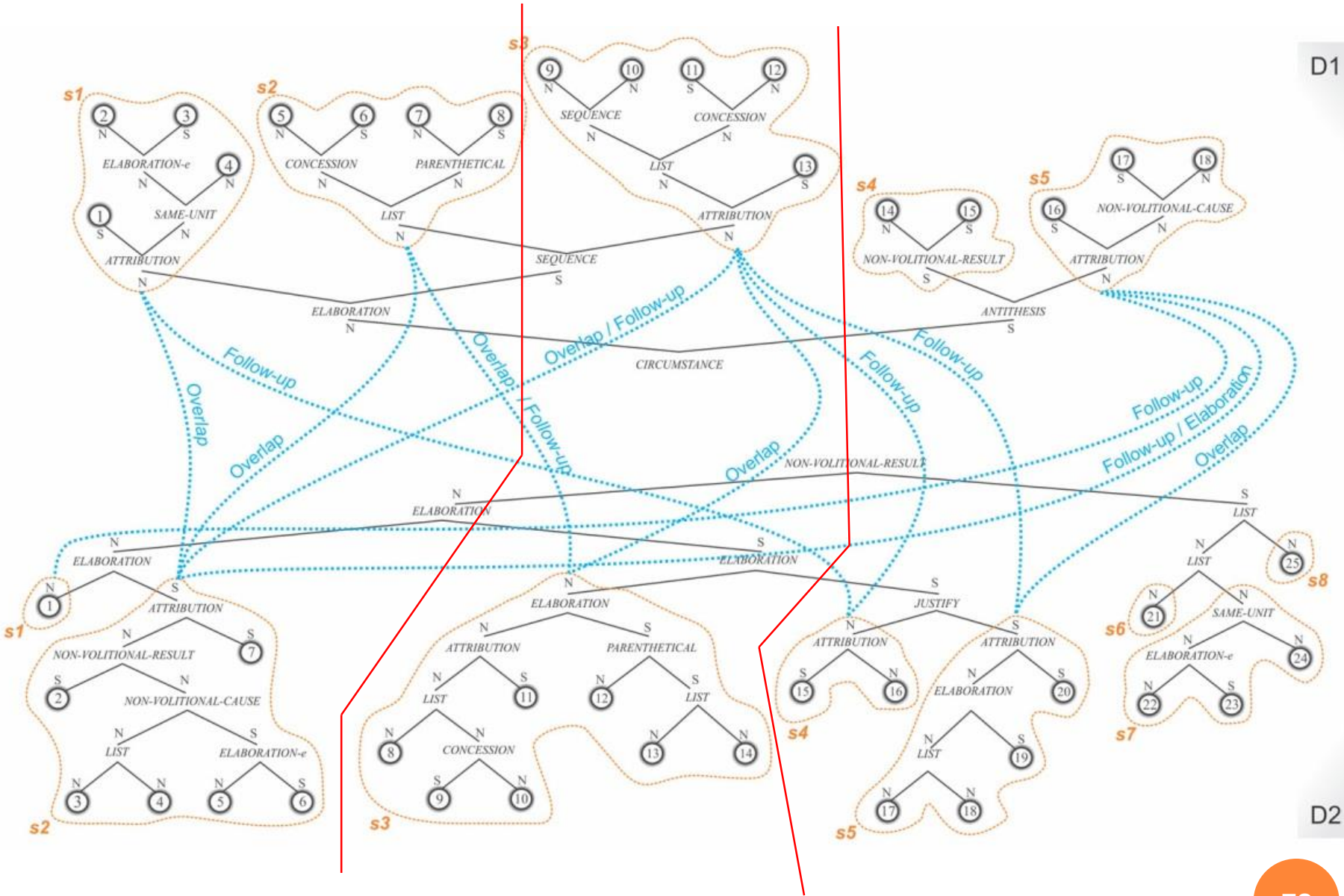
D1

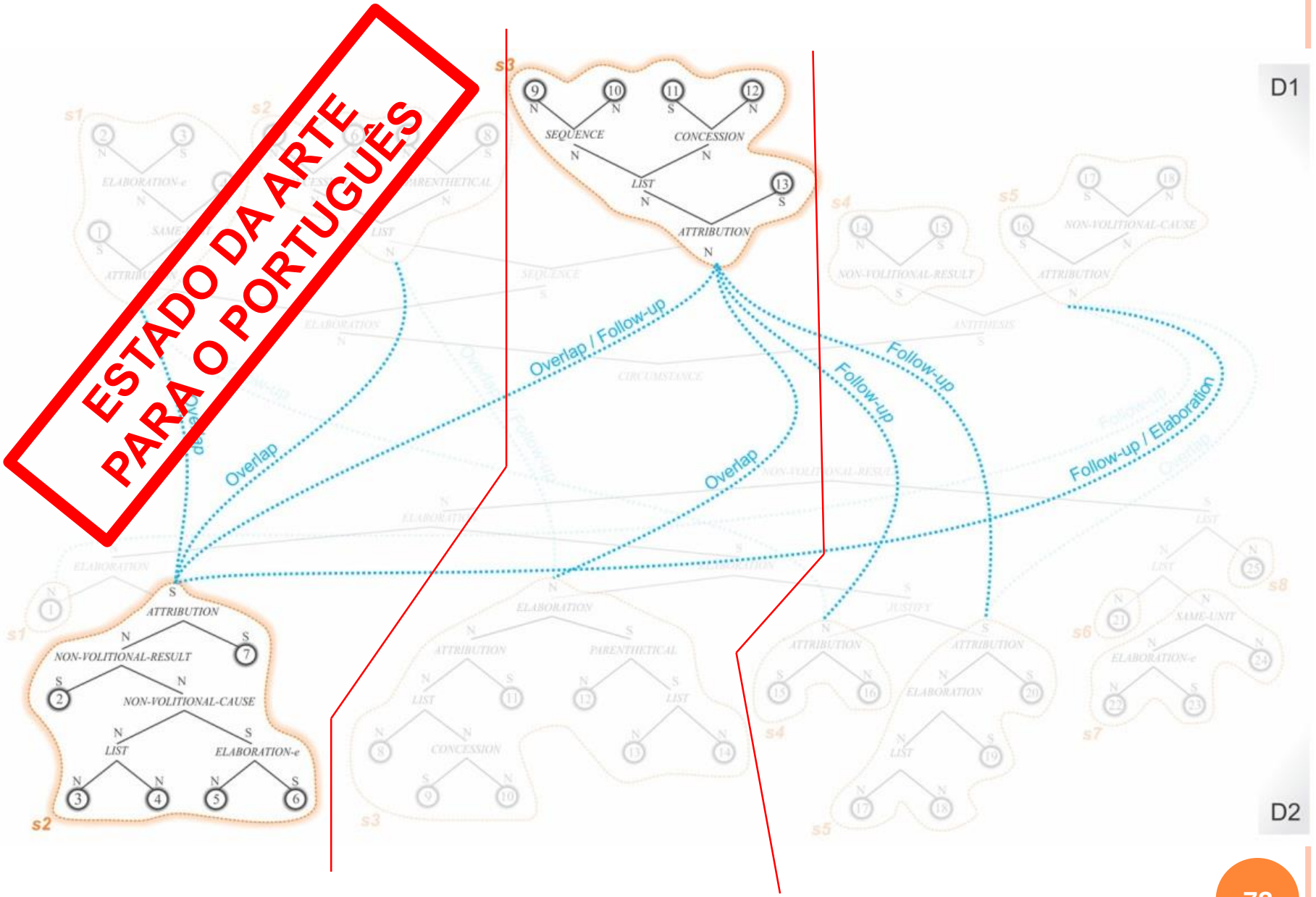

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
ELABORATION-e
SAME-UNIT
CONCESSION
PARENTHETICAL
SEQUENCE
CONCESSION
LIST
ATTRIBUTION
NON-VOLITIONAL-CAUSE
ATTRIBUTION
LIST
SEQUENCE
NON-VOLITIONAL-RESULT
ATTRIBUTION
ELABORATION
ANTITHESIS
CIRCUMSTANCE


NON-VOLITIONAL-RESULT
ELABORATION
LIST
ELABORATION
ELABORATION
LIST
ATTRIBUTION
ELABORATION
JUSTIFY
LIST
SAME-UNIT
NON-VOLITIONAL-RESULT
ATTRIBUTION
PARENTHETICAL
ATTRIBUTION
ATTRIBUTION
ELABORATION-e
NON-VOLITIONAL-CAUSE
LIST
CONCESSION
LIST
ELABORATION
LIST
ELABORATION-e
LIST
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

D2


D1
s1
s2
s3
s4
s5
ELABORATION-e
SAME-UNIT
ATTRIBUTION
CONCESSION
PARENTHETICAL
LIST
SEQUENCE
CONCESSION
LIST
ATTRIBUTION
NON-VOLITIONAL-RESULT
ATTRIBUTION
NON-VOLITIONAL-CAUSE
ELABORATION
ANTITHESIS
CIRCUMSTANCE
Overlap
Follow-up
Overlap / Follow-up
Follow-up
Overlap
Follow-up
Follow-up / Elaboration
Overlap
NON-VOLITIONAL-RESULT
ELABORATION
ELABORATION
LIST
s8
s6
s7
LIST
SAME-UNIT
ELABORATION-e
ATTRIBUTION
NON-VOLITIONAL-RESULT
NON-VOLITIONAL-CAUSE
LIST
ELABORATION-e
ELABORATION
ATTRIBUTION
PARENTHETICAL
LIST
CONCESSION
JUSTIFY
ATTRIBUTION
ATTRIBUTION
ELABORATION
LIST
D2


D1
D2
s1
s2
s3
s4
s5
s6
s7
s8
ELABORATION-e
SAME-UNIT
ATTRIBUTION
CONCESSION
PARENTHETICAL
LIST
SEQUENCE
CONCESSION
LIST
ATTRIBUTION
NON-VOLITIONAL-RESULT
NON-VOLITIONAL-CAUSE
ATTRIBUTION
ANTITHESIS
ELABORATION
CIRCUMSTANCE
Follow-up
Overlap
Overlap / Follow-up
Overlap / Follow-up
Follow-up
Follow-up
Follow-up
Follow-up / Elaboration
Overlap
NON-VOLITIONAL-RESULT
ELABORATION
ELABORATION
LIST
ELABORATION
ELABORATION
JUSTIFY
ATTRIBUTION
NON-VOLITIONAL-RESULT
NON-VOLITIONAL-CAUSE
LIST
ELABORATION-e
ATTRIBUTION
PARENTHETICAL
LIST
CONCESSION
ATTRIBUTION
ATTRIBUTION
ELABORATION
LIST
LIST
SAME-UNIT
ELABORATION-e



Cardoso, P.C.F. and Pardo, T.A.S. (2016). Multi-document summarization using semantic discourse models. Procesamiento de Lenguaje Natural, Vol. 56, pp. 57-64.

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

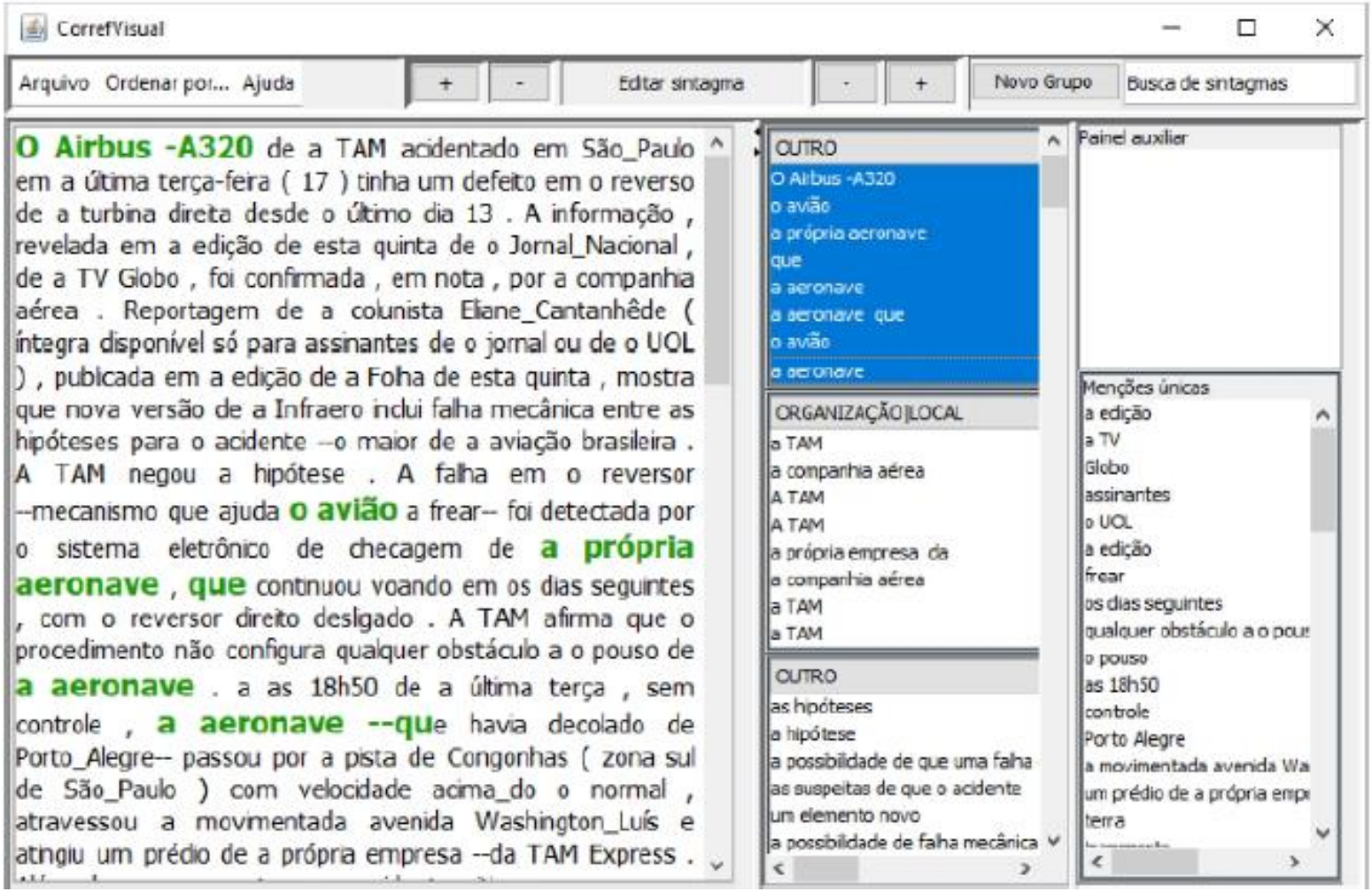
- Exemplo monodocumento

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
Average	*0.54*

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

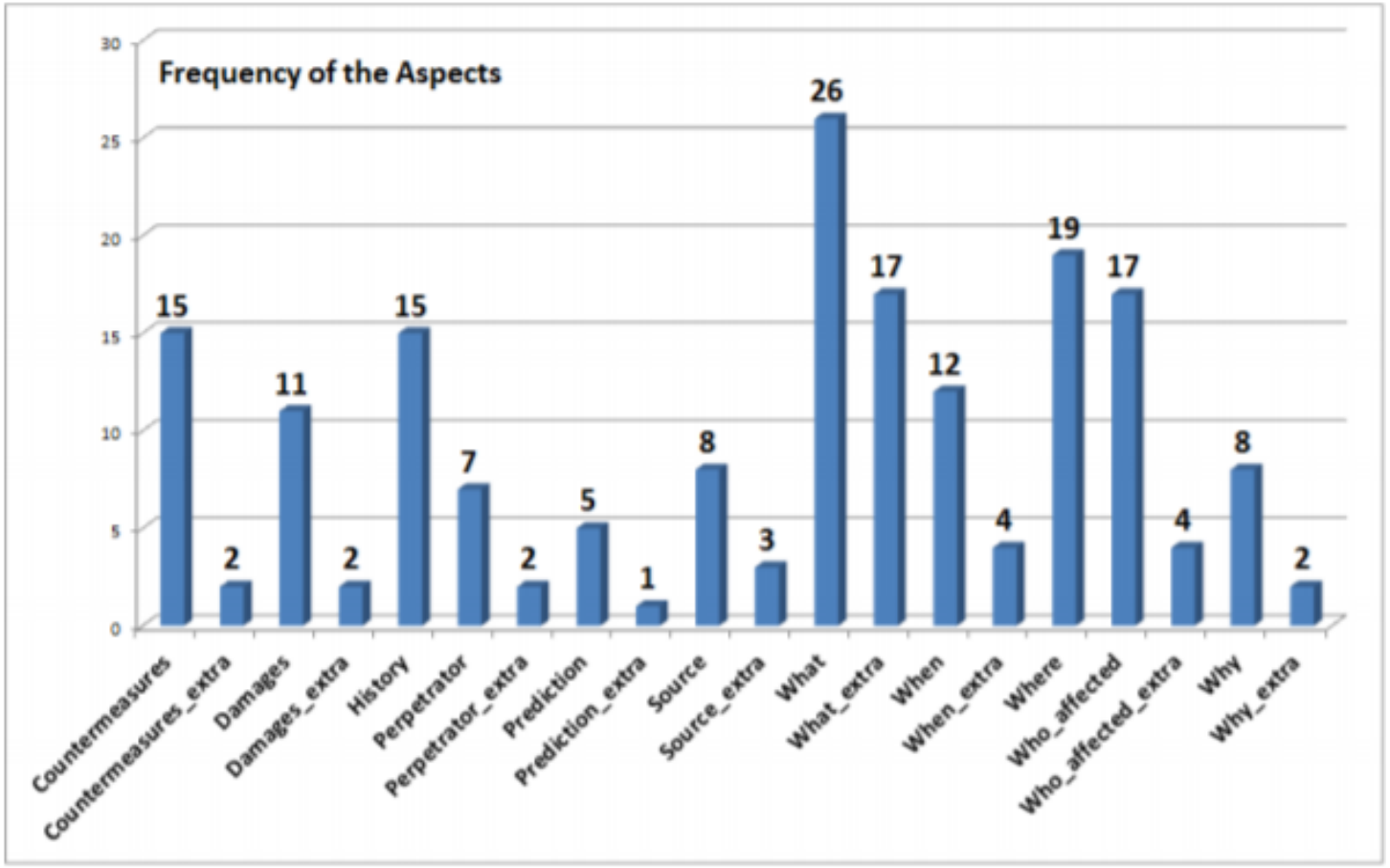
[O avião saiu de Lugushwa a Bukavu e caiu sobre uma floresta]WHERE [após se chocar com uma montanha, prejudicado pelo mau tempo.]WHY

[O avião também levava cargas e minerais.]WHAT_EXTRA

Camargo, R.T.; Maziero, E.G.; Pardo, T.A.S. (2012). Corpus Analysis of Aspects in Multi-Document Summaries - the Case of News Texts from "World" Section. In Anais do XI Encontro de Linguística de Corpus - ELC, pp. 1-10. September 13-15, São Carlos/SP, Brazil.

Aspectos informativos

- Sumários de “mundo”



Aspectos informativos

- Exemplo de padrões “aprendidos” por tópico

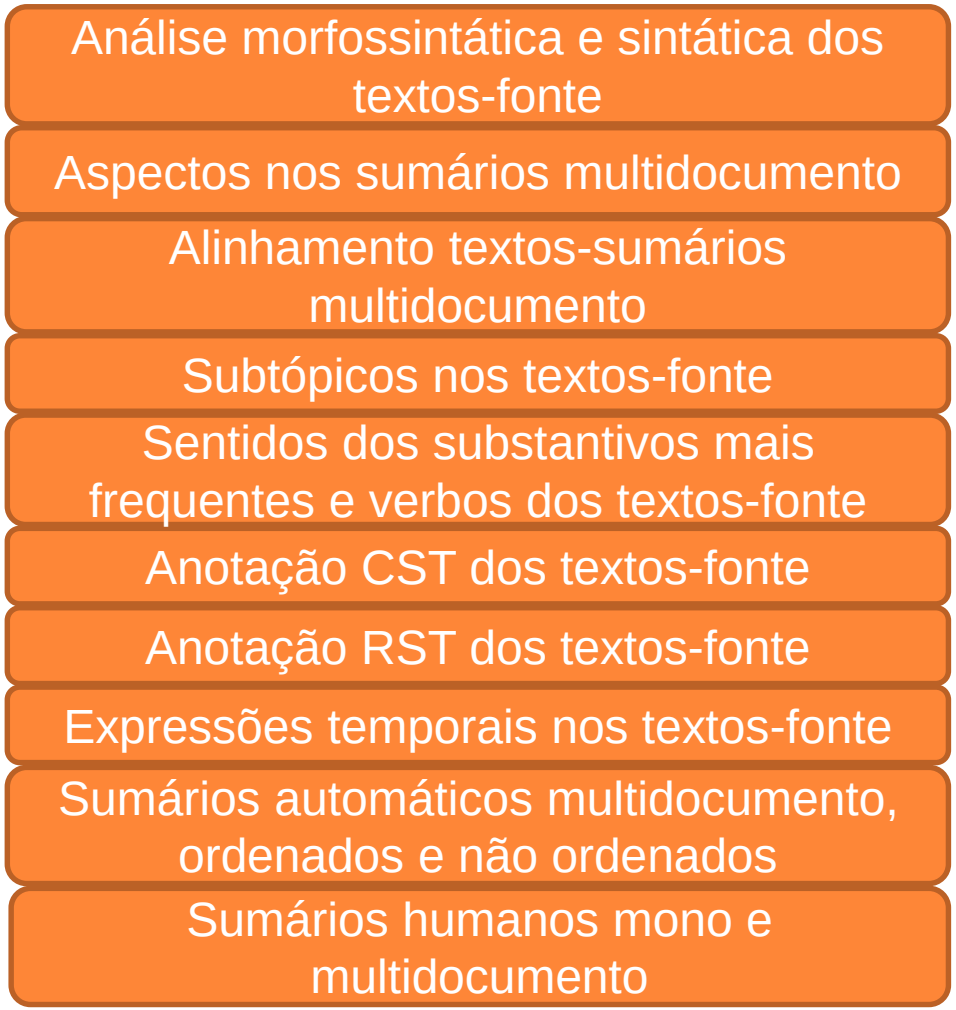
	Accidents	Attacks	Legal and Political Decisions	Natural Disasters
For all summaries				
In common	What, Where, Who_affected, Why	What, Where, Who_affected	What, Perpetrator	What, Where, Who_affected, Countermeasures, Damages
In the 1st paragraph	What, Where, Who_affected	What, Where	What, Perpetrator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



Lições aprendidas

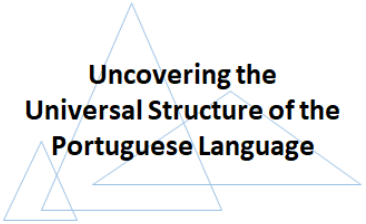
- Anotação de córpus é uma ciência (Hovy e Lavid, 2010)
  - Processo sistemático é necessário
    - Humano: normalmente o mais importante, mas traz grandes desafios!
      - Motivação e grande contribuição
      - Grande aprendizado
      - Afinidade com tarefa, atenção com a língua
      - Projeto da 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



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

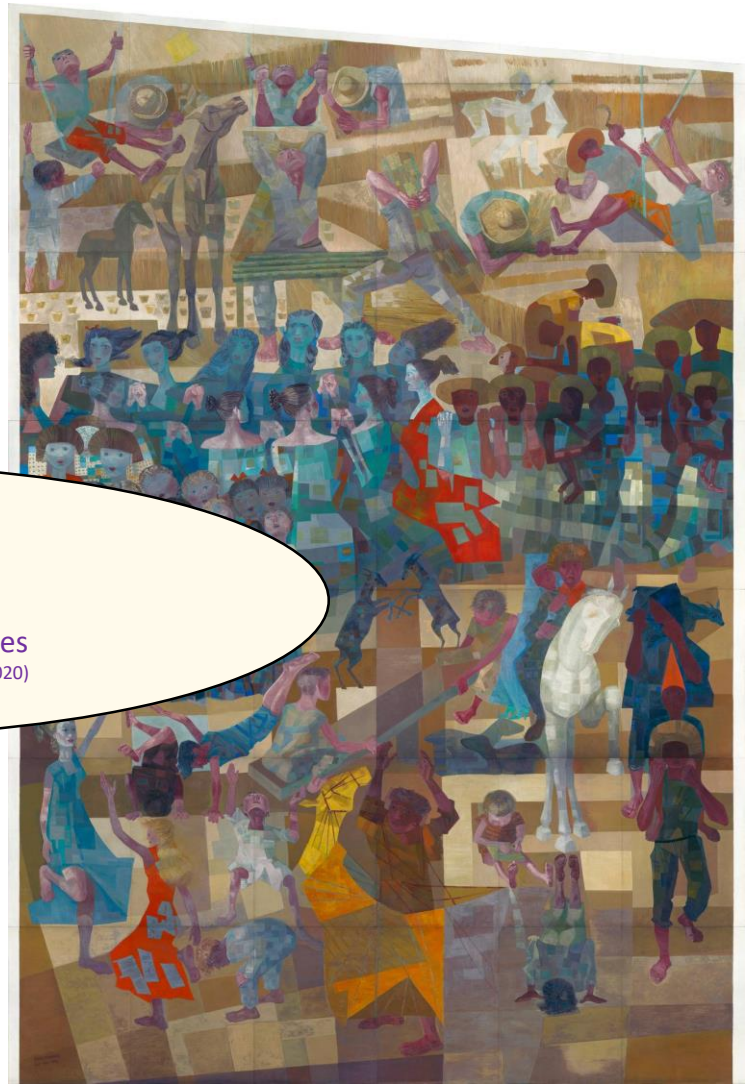
Guerra,
1954

PorTinari

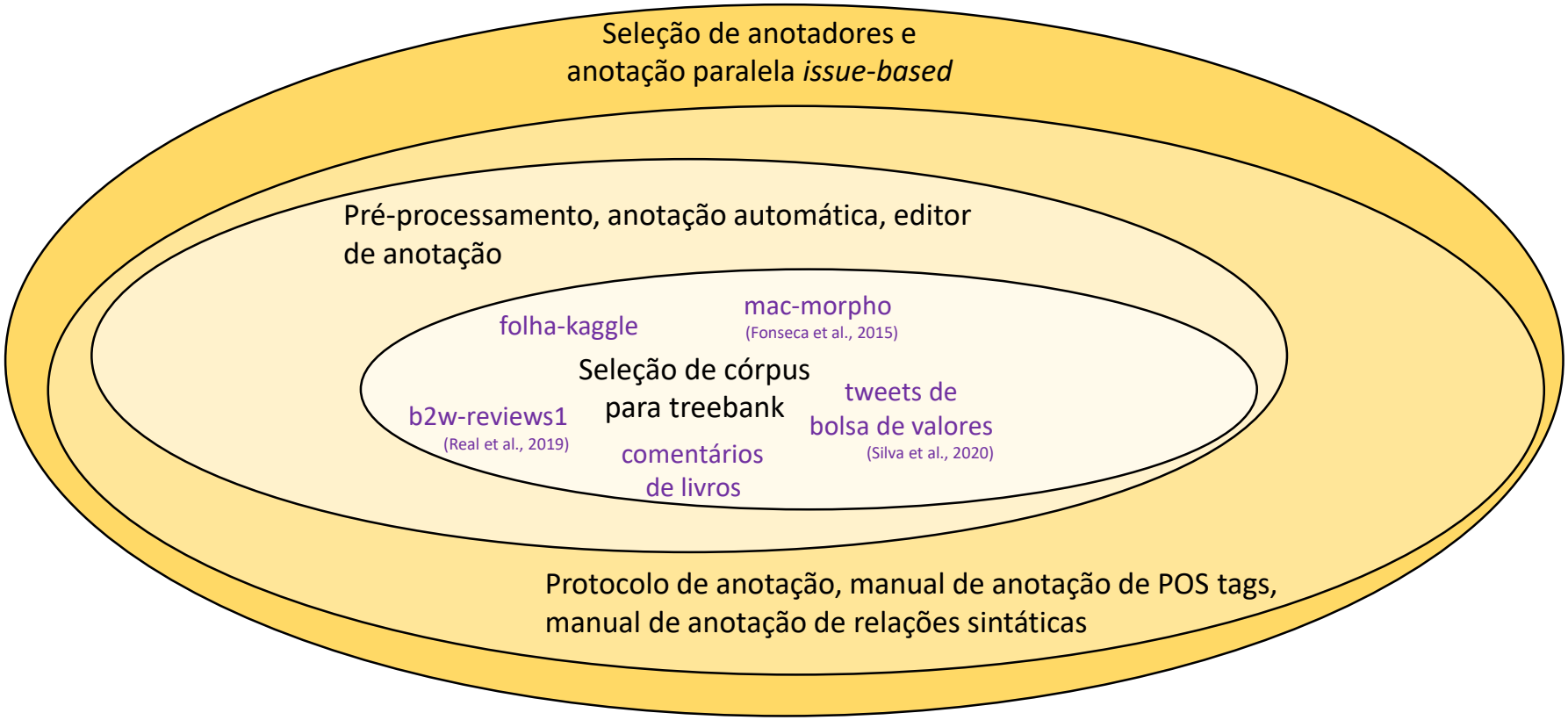
Portuguese Treebank

Seleção de córpus para treebank

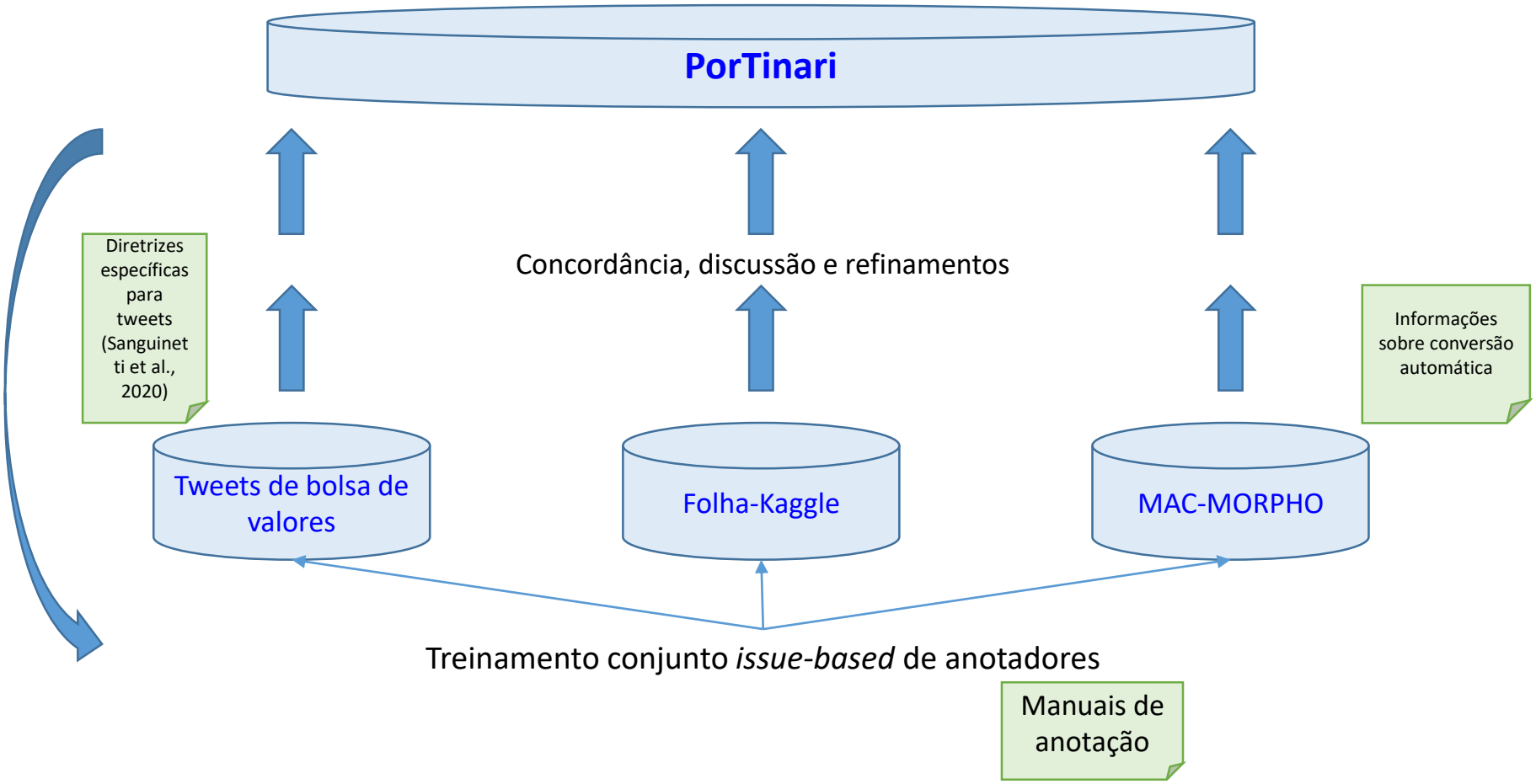
- folha-kaggle
- mac-morpho (Fonseca et al., 2015)
- b2w-reviews1 (Real et al., 2019)
- tweets de bolsa de valores (Silva et al., 2020)
- comentários de livros



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 córpus
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

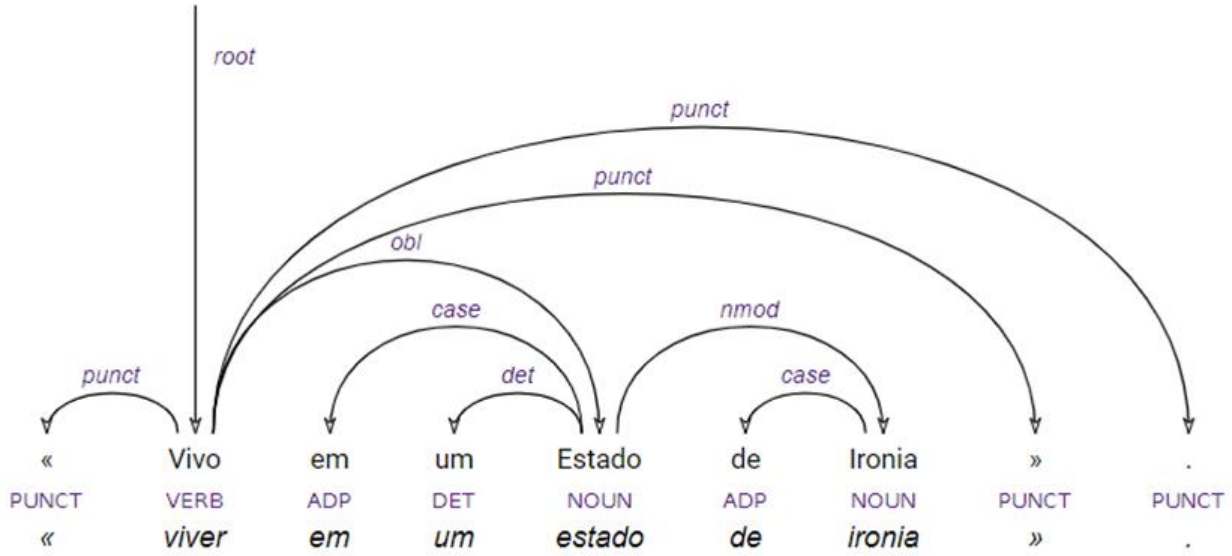
Anotação paralela *issue-based*: o início

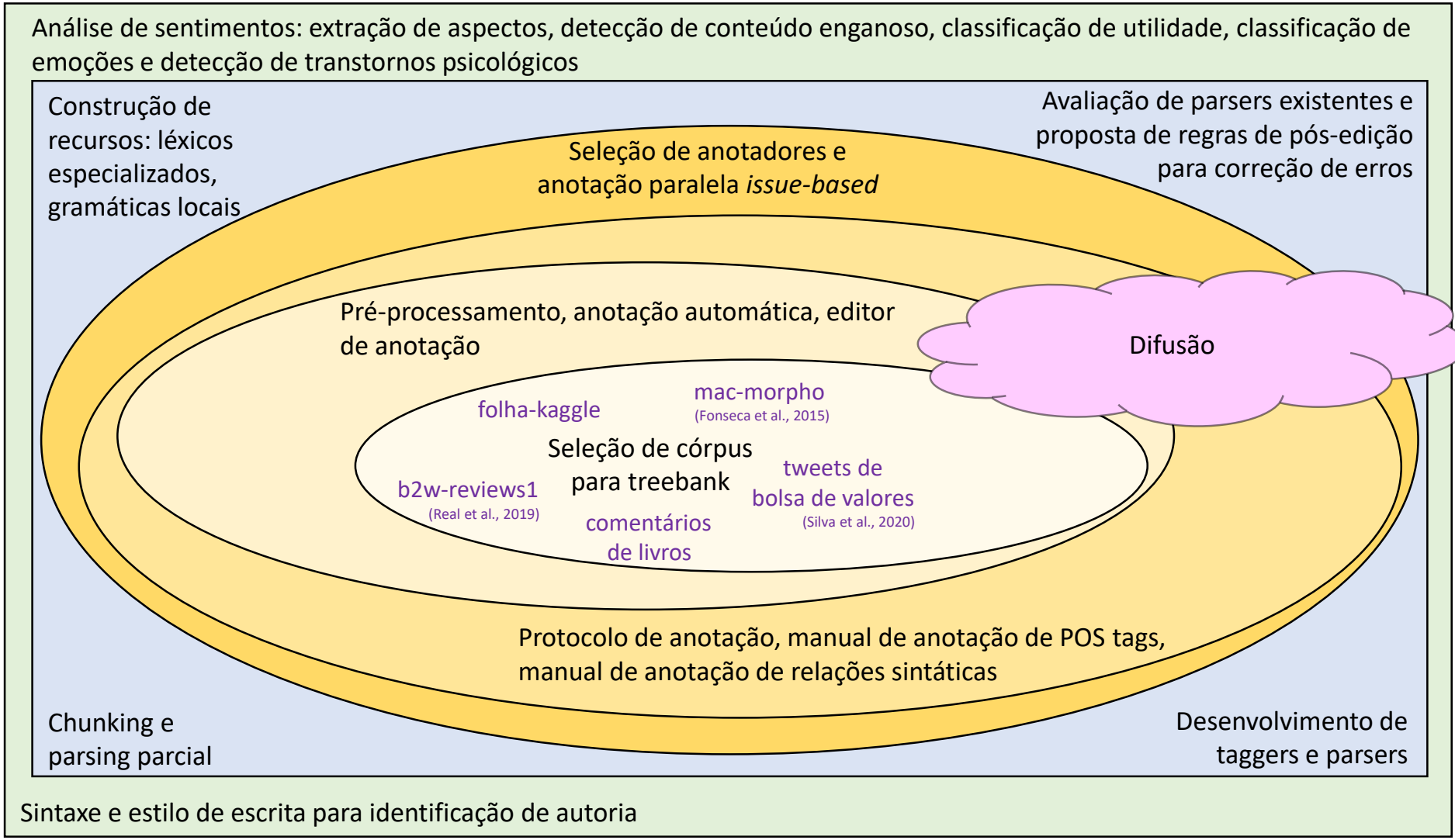


ARBORATOR.grew

Guibon et al. (2020)

Treebank 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 córpus 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