Sumarização Automática Multidocumento

Thiago A. S. Pardo

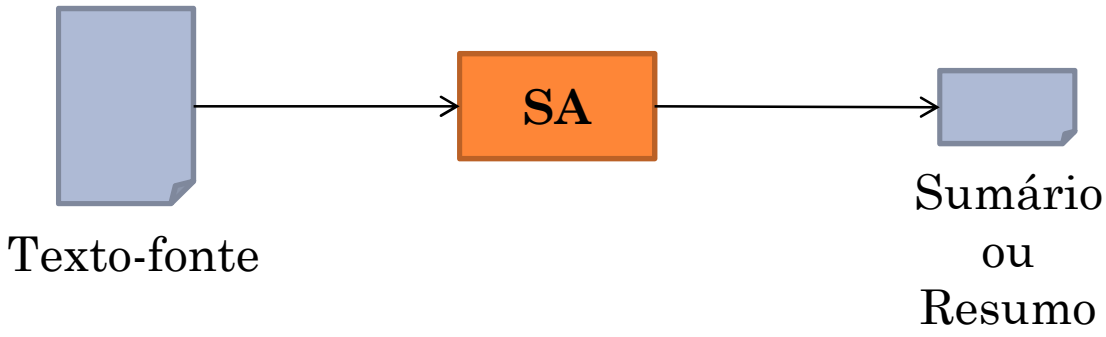
Núcleo Interinstitucional de Lingüística Computacional

Instituto de Ciências Matemáticas e de Computação

Universidade de São Paulo

Sumarização Monodocumento

- Um texto, um resumo



- Estratégia básica: seleção de segmentos textuais mais relevantes, ou omissão de segmentos irrelevantes
  - Primeiros trabalhos "oficiais" datam de 1958

Sumarização Monodocumento

- Definição (Mani, 2001, p. 1)

> *To take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs.*

Sumarização Monodocumento

- Muita história no NILC (Nunes et al., 2010)
  - Modelo profundo de sumarização, combinação de modelos discursivos
    - Rino (1996), Pardo e Rino (2002)
  - Métodos clássicos de sumarização
    - Souza e Nunes (2001), Pereira et al. (2002)
  - Sumarização intrasentencial
    - Martins e Rino (2002), Kawamoto e Pardo (2010)
  - Sistemas simples, extrativos
    - GistSumm (Pardo et al., 2003) e NeuralSumm (Pardo et al. 2003)

SUMARIZAÇÃO MONODOCUMENTO

- Muita história no NILC (Nunes et al., 2010)
  - Estrutura retórica de textos, resolução de correferências
    - Seno e Rino (2005), Carbonel et al. (2007), Tomazela e Rino (2009), Uzêda et al. (2010)
  - Melhores sistemas para o português, usando aprendizado de máquina e redes complexas
    - SuPor e SuPor-2 (Leite et al., 2007, 2008)
  - Pós-edição de sumários para resolução de correferências
    - Gonçalvez et al. (2008)
  - Redes complexas
    - Antiqueira et al., (2009)

Sumarização Monodocumento

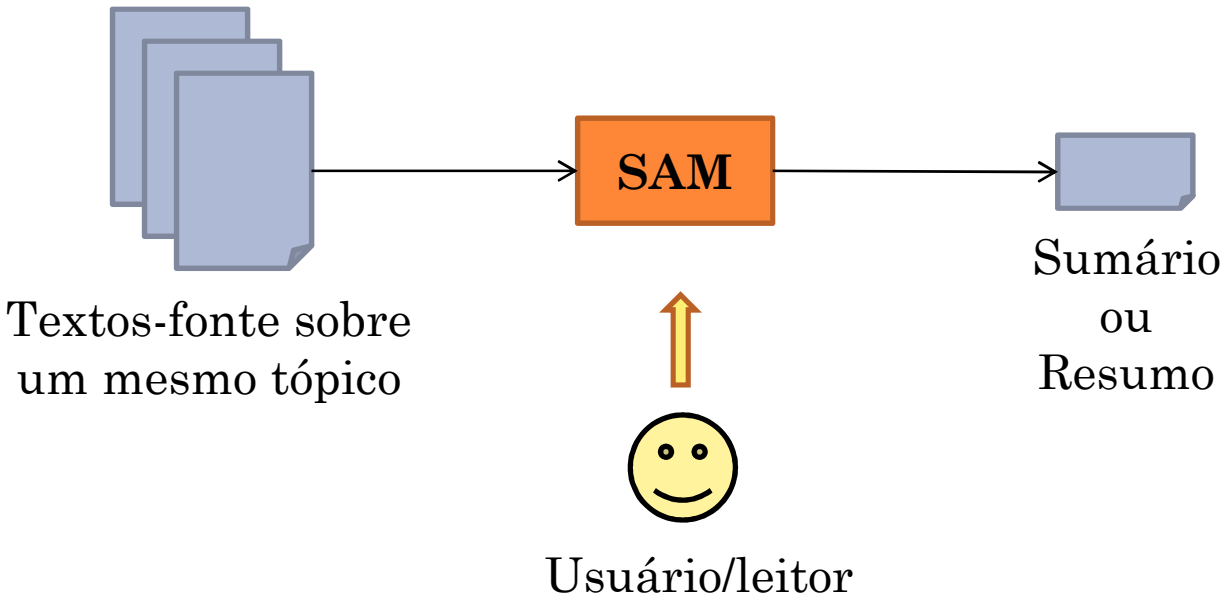
- Muita história no NILC (Nunes et al., 2010)
  - Muitos recursos e ferramentas
    - Córpus
    - Ferramentas de anotação de córpus
    - Ferramentas de processamento de textos e sumários
    - Léxicos especializados
    - Etc.

Sumarização Monodocumento

- Muitos desafios
  - Manutenção da informatividade
  - Coerência e coesão dos textos
    - Resolução de expressões referenciais, elementos textuais de coesão, estrutura temática, etc.
  - Clareza e fluência textual
  - Manutenção da taxa de compressão
  - *Abstracts*, sumários críticos
  - Resultados praticamente estagnados

SUMARIZAÇÃO MULTIDOCUMENTO

- Sumarização Automática Multidocumento (SAM)



SUMARIZAÇÃO MULTIDOCUMENTO

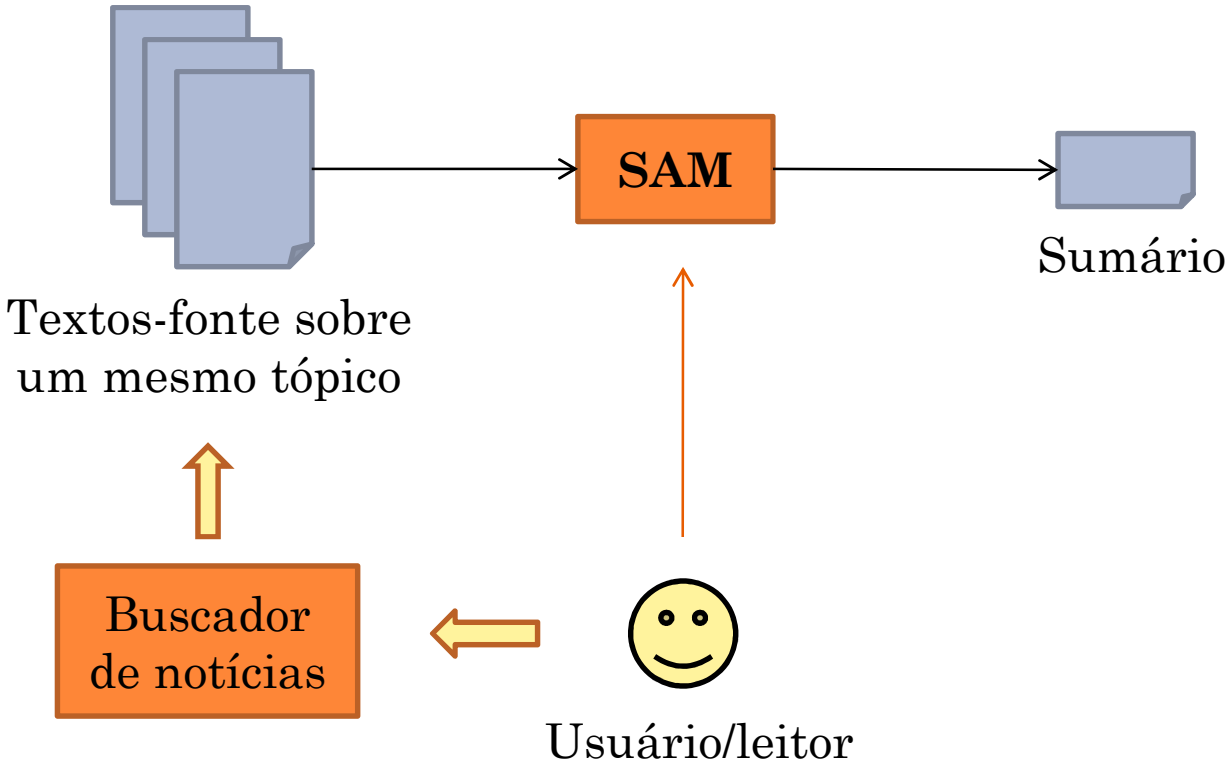
- Definição (Mani, 2001, p. 169)

> *To take an information source that is a collection of related documents and extract content from it, while removing redundancy and taking into account similarities and differences in information content, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs.*

SUMARIZAÇÃO MULTIDOCUMENTO

- Prefácio de livros
- Texto introdutório de revistas/coletâneas de artigos
- Índices de livros e revistas temáticas
- Biografias
- Organização de notícias e links em portais web
  - **Web!**

CENÁRIO MAIS ATUAL



SUMARIZAÇÃO MULTIDOCUMENTO

- História relativamente recente
  - McKeown e Radev, 1995
- Sumarização monodocumento vs. multidocumento (Mani, 2001)
  - Mesmos problemas originais, novos desafios
- Estimativas do IDC: 800 exabytes de informação nova em 2009

Exemplo

- Radev e McKeown (1998, p. 478)
  - Sumário de 4 textos sobre terrorismo

Reuters reported that 18 people were killed in a Jerusalem bombing Sunday. The next day, a bomb in Tel Aviv killed at least 10 people and wounded 30 according to Israel radio. Reuters reported that at least 12 people were killed and 105 wounded. Later the same day, Reuters reported that the radical Muslim group Hamas had claimed responsibility for the act.

chile - Google Notícias

http://news.google.com.br/news/search?pz=1&cf=all&ned=pt-BR_br&hl=pt-BR&q=chile

Sumarizar

Web Imagens Vídeos Mapas **Notícias** Orkut Gmail mais ▼ Fazer login

Google notícias Brasil

chile Pesquisar Pesquisa avançada de notícias Preferências

Notícias Resultados **1 – 10** de aproximadamente **178.465** para **chile**. (**0,14** segundos)

› **Qualquer conteúdo**
Imagens
Blogs

› **Quaisquer notícias recentes**
Última hora
Dia anterior
Última semana
Último mês
2010
2009
2008
2007
2006
Arquivos

› **Classificado por relevância**
Classificado por data


bbc brasil

Bachelet mantém popularidade de 84% após tremor no **Chile**

O Globo - há 2 horas

Uma pesquisa de opinião divulgada nesta terça-feira indica que a presidente do **Chile**, Michelle Bachelet, manteve 84% de aprovação popular após o terremoto ...

Popularidade de Bachelet no **Chile** se mantém intacta após terremoto AFP
Novo presidente do **Chile** prepara plano financeiro e acordos DCI
Aprovação de Bachelet se mantém em 84% após terremoto G1.com.br
Monitor Mercantil - Guia Global
todos os 76 artigos » Enviar por e-mail


Band

Terremoto no **Chile** alterou posição da cidade de Concepción em três ...

O Globo - há 10 minutos

RIO - A magnitude do terremoto que atingiu o **Chile** há dez dias foi violenta o suficiente para alterar a posição relativa da cidade de Concepción pelo menos ...

Terremoto desloca cidades no **Chile** O São Gonçalo
Estudo conjunto de universidades Correio da Manhã
Terremoto deslocou cidade chilena AFP
JC OnLine - Fábrica de Conteúdos
todos os 35 artigos » Enviar por e-mail


Estadão

Chile confirma identidade de 497 mortos em terremoto

Estadão - há 17 horas

Também nesta segunda-feira, o presidente eleito do **Chile**, Sebastián Piñera, que tomará posse na quinta-feira, disse que os militares enviados à costa oeste ...

Mortos identificados no terremoto do **Chile** são 497 Abril
Chile identifica 497 mortos em terremoto e tsunami AFP
Sobe para 497 número de mortos identificados após terremoto no **Chile** SRZD
Zero Hora - Band
todos os 208 artigos » Enviar por e-mail

Forças Armadas dos EUA enviam médicos para o **Chile**

G1.com.br - há 1 hora

As Forças Armadas americanas anunciaram nesta terça-feira (9) o envio de dezenas de médicos para o **Chile** para fornecer assistência aos sobreviventes do ...

DESAFIOS MULTIDOCUMENTO

- Conceitos importantes
  - Taxa de compressão
  - Audiência: sumários genéricos ou focados nos interesses dos usuários
  - Extratos vs. *abstracts*
  - Sumários informativos, indicativos ou críticos
  - Texto vs. fragmentos

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 "lingüístico-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 "lingüístico-computacionais"
  - 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

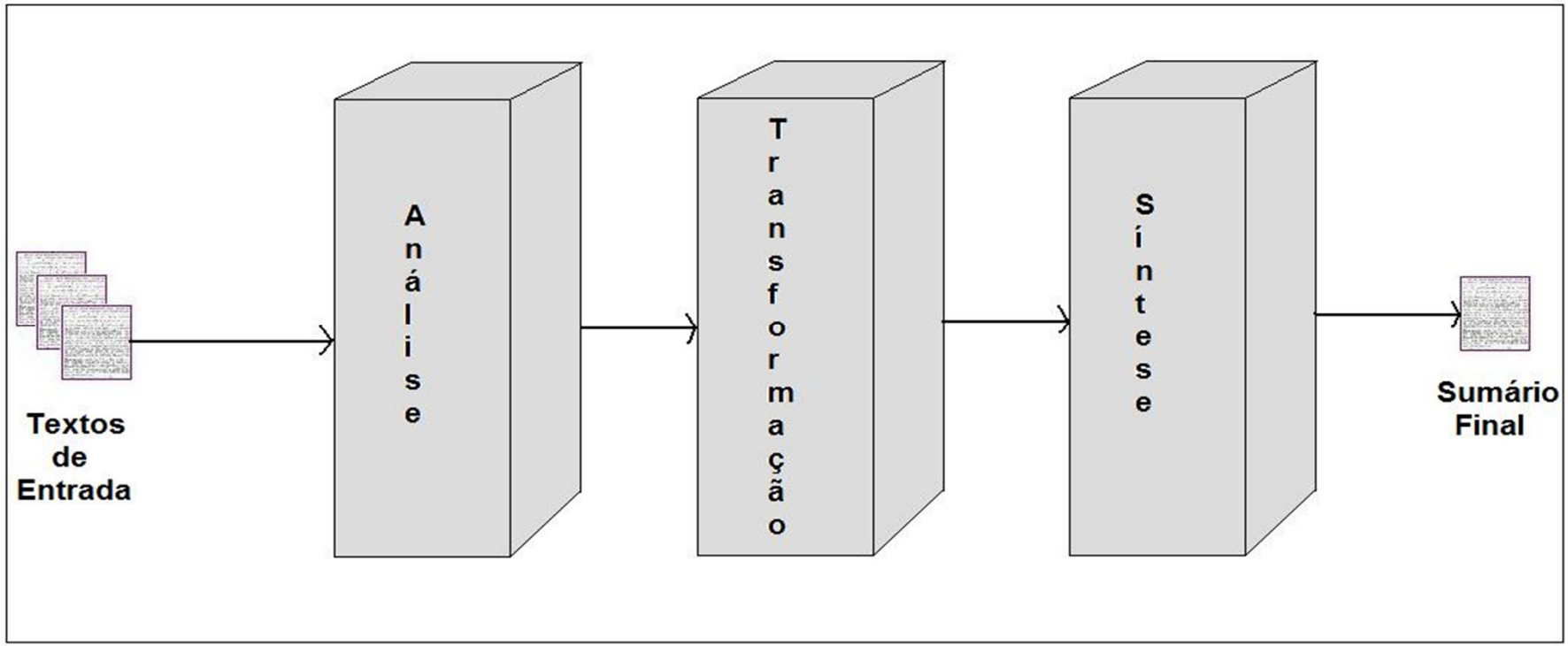
DESAFIOS MULTIDOCUMENTO

- 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

VANTAGENS MULTIDOCUMENTO

- A informação importante se repete ou é elaborada nos vários textos
  - Muito mais evidência do que na sumarização monodocumento
    - Redundância pode ser útil!
      - Formas de ocorrência
        - Equivalências exatas ou parciais, paráfrases, subsunções, resumos

Arquitetura Genérica de SA



SUMARIZAÇÃO MULTIDOCUMENTO

- Humanos
  - Tarefa não natural
  - ... mas útil, como experimentos já constataram
  - Indícios da tarefa humana
    - Primeiro se seleciona um texto-fonte base
    - Extraem-se do texto-fonte base a informação principal inicial
    - Complementa-se a informação extraída com informações dos outros textos
      - Escolha do texto-fonte base depende de vários fatores: conteúdo, data de publicação, fonte, demanda, tempo disponível

Abordagens ao problema

- Superficial
  - Pouco uso de conhecimento lingüístico
    - Empirismo e estatística (básica ou avançada)
  - Mais barato, mas com resultados piores
- Profunda
  - Uso de mais conhecimento
    - Sintaxe, semântica, discurso e pragmática
  - Mais caro, mas com resultados melhores
- Híbrida
  - O melhor dos dois mundos
    - Mas como unificar os dois mundos?

Abordagens ao problema

- Superficial
  - Pouco uso de conhecimento lingüístico
    - Empirismo e estatística (básica ou avançada)
  - Mais barato, mas com resultados piores

- Profunda
  - Uso de mais conhecimento
    - Sintaxe, semântica, discurso e pragmática
  - Mais caro, mas com resultados melhores

- Híbrida
  - O melhor dos dois mundos
    - Mas como unificar os dois mundos?

ABORDAGENS SUPERFICIAIS

- **Métodos a la Salton (1997)**
  - Ando et al. (2000), Mihalcea e Tarau (2005)
    - Texto é segmentado
      - Palavras, orações, sentenças, parágrafos ou blocos maiores
    - Medidas de similaridade lexical são aplicadas entre os segmentos, estabelecendo relações mais fortes ou mais fracas entre eles
      - Dice, Jaccard, Cosseno, simples co-ocorrência de palavras, etc.
        - Representação via grafos é comum
    - Melhores são selecionados para o sumário

ABORDAGENS SUPERFICIAIS

- **Métodos a la Salton (1997)**
  - Ando et al. (2000), Mihalcea e Tarau (2005)
    - Texto é segmentado
      - Palavras, orações, sentenças, parágrafos ou blocos maiores
    - Medidas de similaridade lexical são aplicadas entre os segmentos, estabelecendo relações mais fortes ou mais fracas entre eles
      - Dice, Jaccard, Cosseno, simples co-ocorrência de palavras, etc.
    - Me

Métodos robustos, mas com sérios problemas: não se trata redundância, termos sinônimos ou correlatos normalmente não são tratados

ABORDAGENS SUPERFICIAIS

- Relevância vs. redundância
  - *Maximal Marginal Relevance* – MMR (Carbonel et al., 1997)
    - Nota de um segmento = relevância – redundância, podendo haver pesos diferentes
      - Focado ou não nos interesses do usuário
        - Se focado, a consulta (*query*) é importante
  - Goldstain (2000)
    - Melhoria da MMR com atributos de posição dos segmentos, presença de entidades nomeadas, presença de termos da *query*, data de publicação do documento, etc.
  - *Interactive* MMR (Lin et al. 2010)
    - Sumário construído colaborativamente com o usuário, que escolhe a melhor sentença dentre o ranque formado pelo MMR

ABORDAGENS SUPERFICIAIS

- Combinação de atributos textuais
  - Radev et al. (2000, 2001) e o sistema MEAD
    - Segmentos são julgados em função de

      (i) distância lexical em relação ao centróide dos documentos
      (ii) distância em relação ao início dos textos a que pertencem
      (iii) distância lexical em relação à primeira sentença dos textos a que pertencem

    - Verificação de redundância entre segmentos que vão para o sumário

ABORDAGENS SUPERFICIAIS

- Relações lexicais
  - Mani e Bloedorn (1999)
    - Relações de proximidade, correferência, sinonímia e hiperonímia
    - Grafos de relacionamentos entre termos

ABORDAGENS SUPERFICIAIS/PROFUNDAS

- Sintaxe
  - Barziley et al. (1999)
    - Procuram-se equivalências entre segmentos pela comparação entre seus objetos e ações
    - Segmentos identificados formam grupos temáticos
    - Segmentos de cada grupo temático são selecionados para o sumário (via algum dos métodos anteriores)
    - Operações de intersecção de segmentos também são realizadas com base em sintaxe

Abordagens ao problema

- Superficial
  - Pouco uso de conhecimento lingüístico
    - Empirismo e estatística (básica ou avançada)
  - Mais barato, mas com resultados piores

- Profunda
  - Uso de mais conhecimento
    - Sintaxe, semântica, discurso e pragmática
  - Mais caro, mas com resultados melhores

- Híbrida
  - O melhor dos dois mundos
    - Mas como unificar os dois mundos?

ABORDAGENS PROFUNDAS

- Relações semânticas entre mensagens
  - McKeown e Radev (1995)
    - Textos sobre terrorismo são interpretados e representados em templates
      - Perpetrador, vítimas, tipo de evento, local, data, etc.
        - Uso de ferramentas de extração de informação (provenientes da MUC)
    - Templates similares são fundidos via operadores, gerando templates mais relevantes
      - Operadores de contradição, refinamento,/especialização, mudança de perspectiva, etc.

ABORDAGENS PROFUNDAS

- Relações discursivas entre textos
  - Radev (2000)
    - Trabalho teórico, principal representante da linha profunda
    - *Cross-document Structure Theory* (CST)
      - Por exemplo, relações de contradição, redundância, complementaridade, estilo, etc.
    - Operadores de sumarização com base nas relações e nas preferências de usuários

ABORDAGENS PROFUNDAS

- Enriquecimento de sumários
  - Zhang et al. (2002)
    - Trocam-se sentenças do sumário por sentenças mais importantes
      - Sentenças retiradas: baixa pontuação e sem/poucas relações CST
      - Sentenças adicionadas: com relações CST com outras sentenças no sumário
        - Verificação: sumários cujas sentenças são mais relacionadas (pelas relações CST) são melhores

ABORDAGENS PROFUNDAS

- Relações discursivas de domínio
  - Afantenos et al. (2004, 2007) e proposta teórica
    - Relações sincrônicas e diacrônicas para um domínio
    - Relacionamento entre mensagens (com indexação ontológica)
      - Necessidade de extração de informação
    - Tratamento de mensagens para compor o sumário em função das relações entre elas

Abordagens ao problema

- Superficial
  - Pouco uso de conhecimento lingüístico
    - Empirismo e estatística (básica ou avançada)
  - Mais barato, mas com resultados piores

- Profunda: foco do único projeto no Brasil
  - Uso de mais conhecimento
    - Sintaxe, **semântica**, **discurso** e pragmática
  - Mais caro, mas com resultados melhores

- Híbrida
  - O melhor dos dois mundos
    - Mas como unificar os dois mundos?

ABORDAGENS AO PROBLEMA

- Superficial
  - P
- Profu
  - Uso de mais conhecimento
    - Sintaxe, **semântica**, **discurso** e pragmática
  - Mais caro, mas com resultados melhores
- Híbrida
  - O melhor dos dois mundos
    - Mas como unificar os dois mundos?

Hipótese de pesquisa: é necessário conhecimento mais sofisticado para lidar adequadamente com os fenômenos multidocumento

Um pouco de história

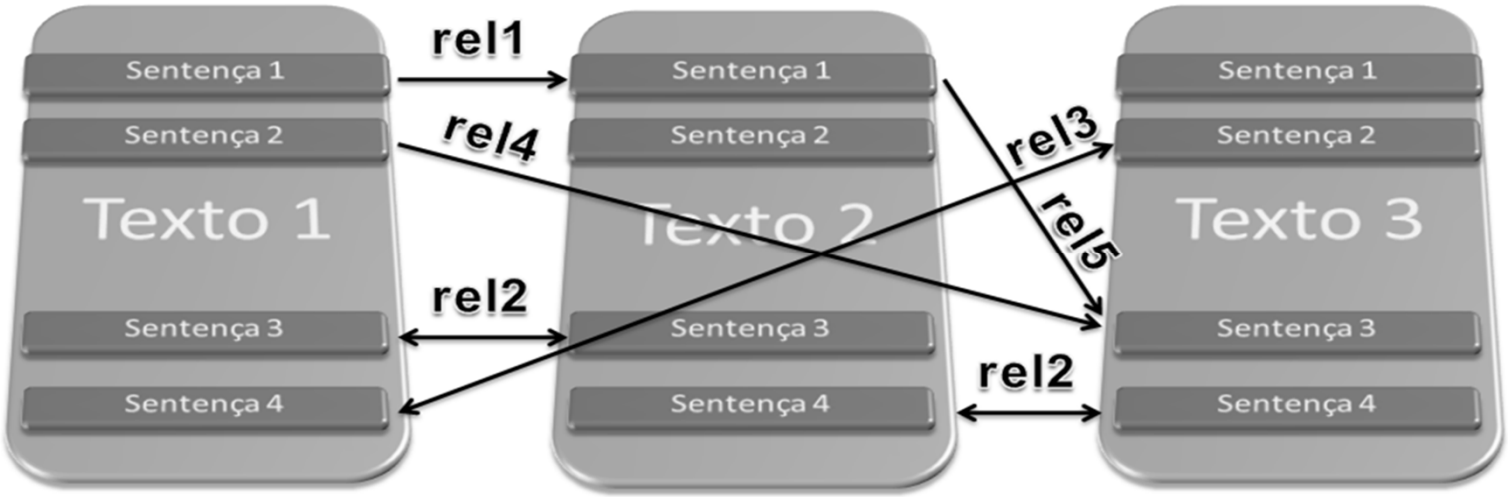
- Trigg e o sistema TextNet (1983, 1986)
- RST (Mann e Thompson, 1987)
- Radev e Mckeown (1995): SUMMONS e seus operadores
- Radev (2000): CST (*Cross-document Structure Theory*)
- Afantenos et al. (2004) e críticas a CST
- Sucesso em aplicações de sumarização multidocumento (Radev et al., 2000, 2001; Zhang et al., 2002; Afantenos et al., 2004, 2007)

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

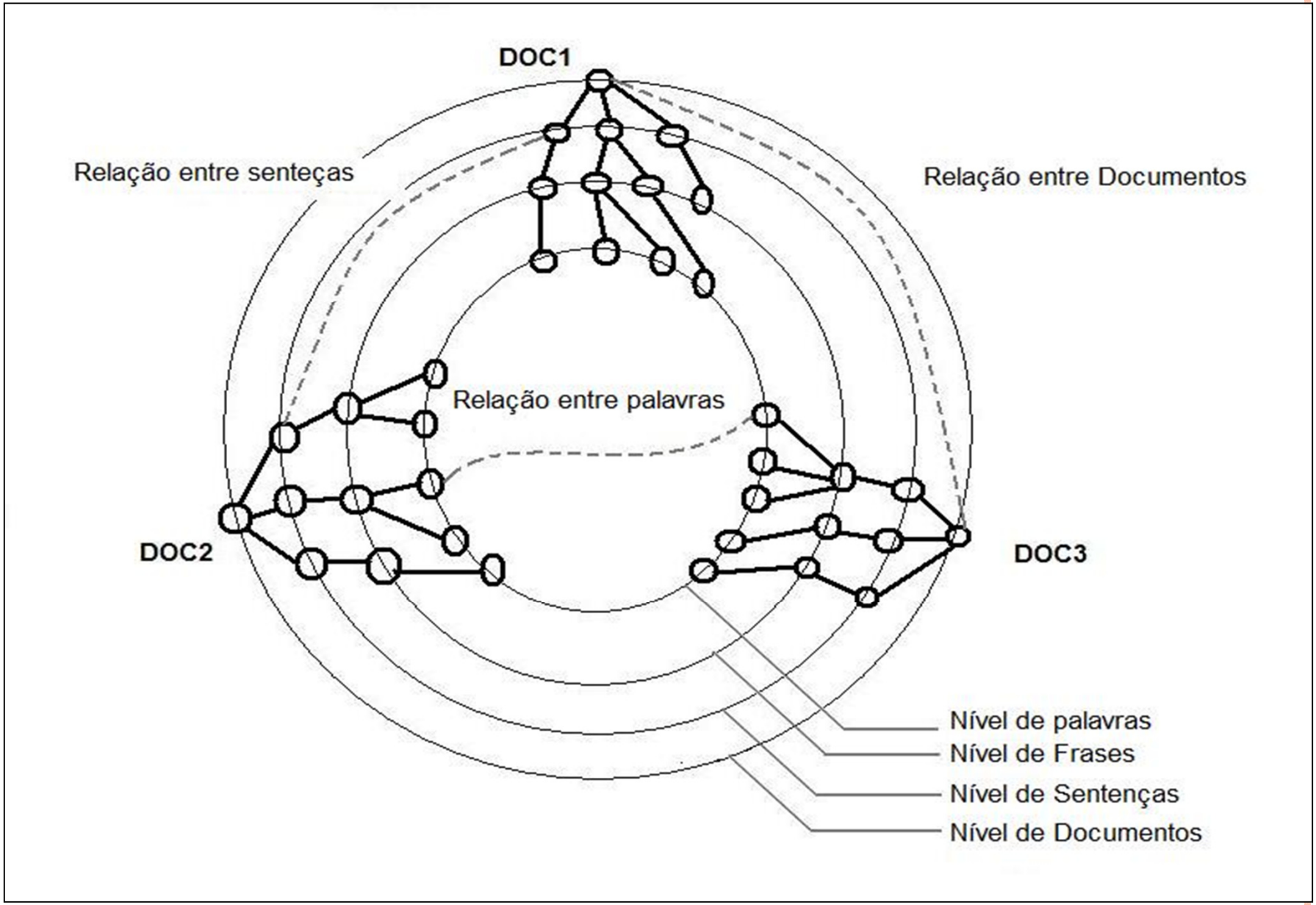
- Teoria discursiva multidocumento
- 24 relações para documentos que versam sobre um mesmo assunto
- Estruturas de dados complementares
  - Cubo multidocumento: fonte, tempo e posição dos segmentos textuais
  - Grafo multidocumento: relações multidocumento

CST

- Modelo semântico-discursivo de estruturação multidocumento
  - São definidas relações entre partes dos documentos/textos

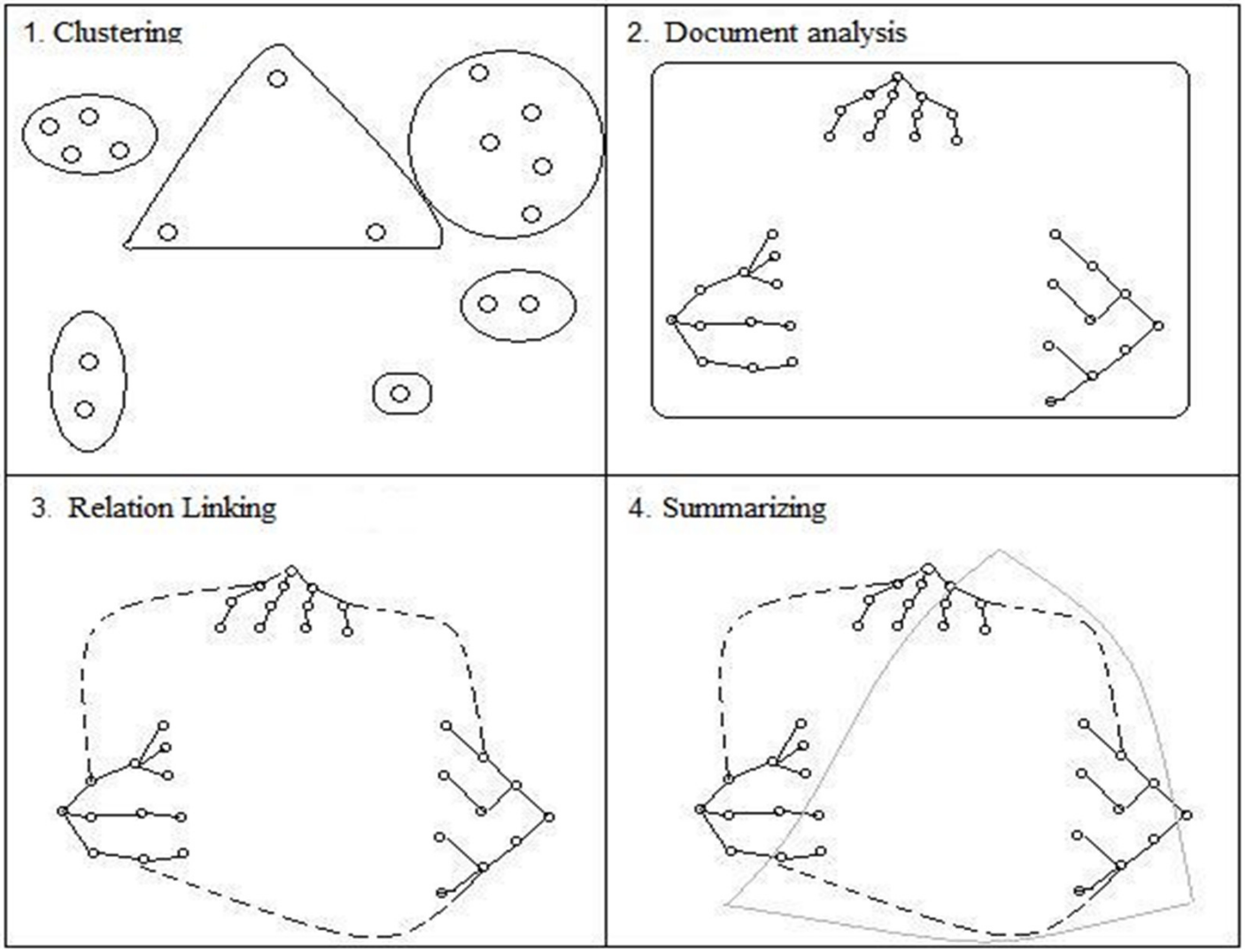


CST


DOC1
Relação entre senteças
Relação entre Documentos
Relação entre palavras
DOC2
DOC3
Nível de palavras
Nível de Frases
Nível de Sentenças
Nível de Documentos

CST e sumarização

- Radev (2000): 4 etapas



CST para o inglês

- CSTBank (Radev, 2003)
  - Baixa concordância entre anotadores
  - *Since it describes relationships that hold across multiple documents rather than across spans of text within the same document, it makes no assumptions about authors' intentions in creating cohesion in texts*

CST

- Relações originais

Identity	Modality	Judgment
Equivalence	Attribution	Fulfillment
Translation	Summary	Description
Subsumption	Follow-up	Reader profile
Contradiction	Elaboration	Contrast
Historical background	Indirect speech	Parallel
Cross-reference	Refinement	Generalization
Citation	Agreement	Change of perspective

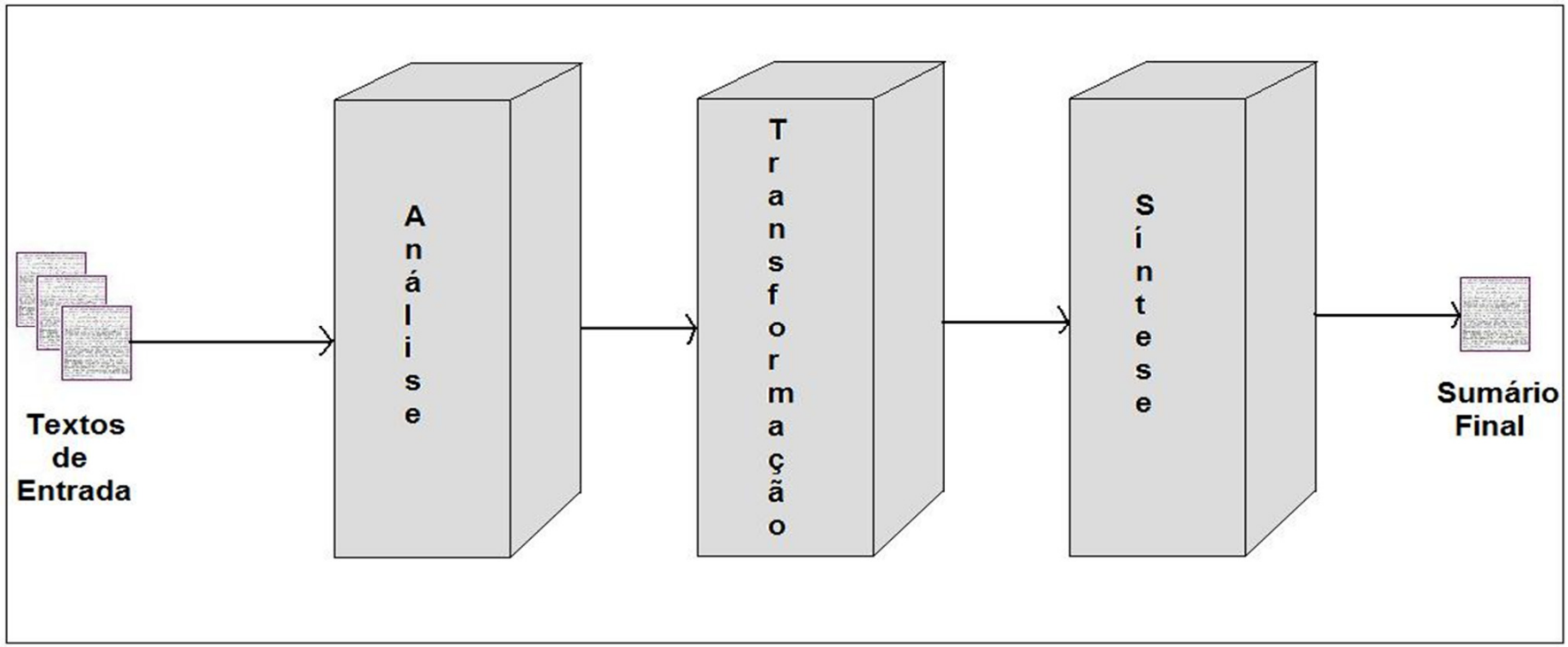
Exemplo

- *Contradiction*, *overlap*, *historical background* (←)

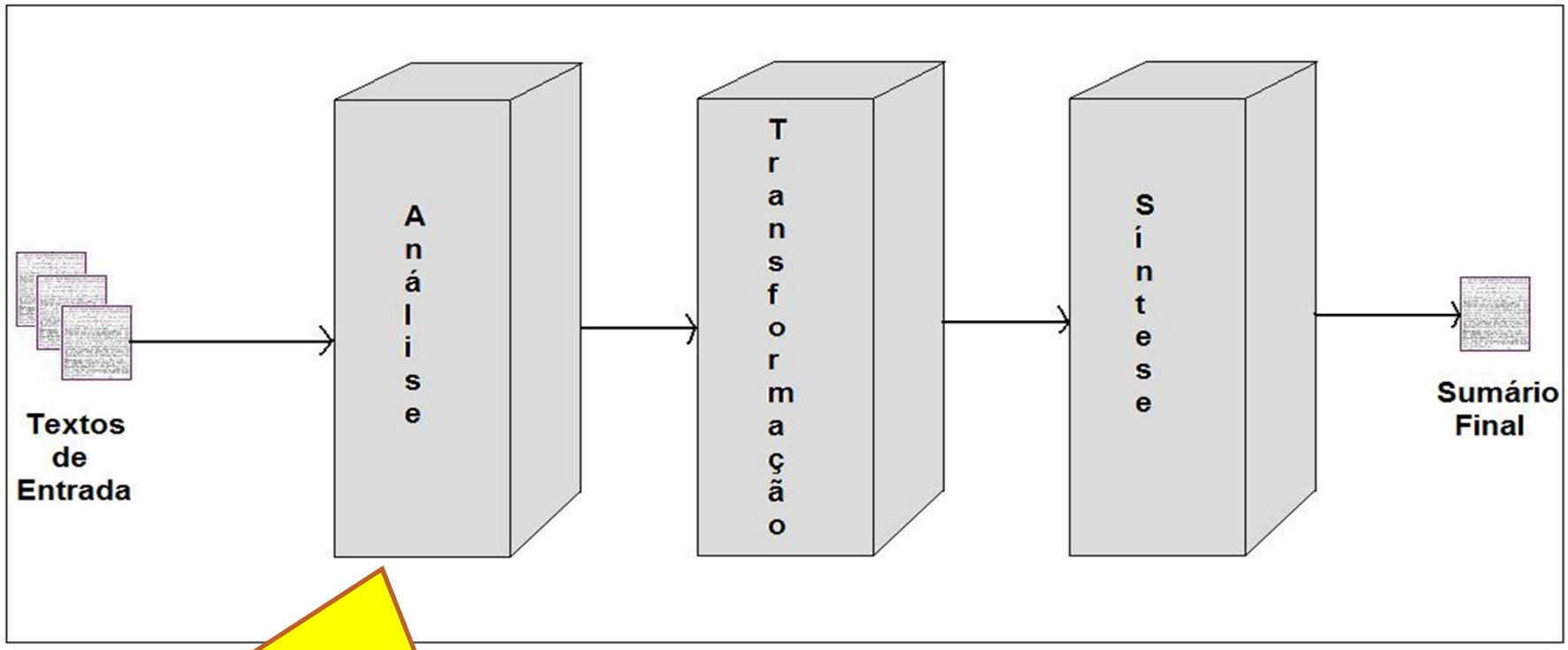
D1: Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 13 pessoas na quinta-feira à tarde, informou nesta sexta feira um porta-voz das Nações Unidas.

D2: Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. Segundo um porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade. O Congo tem um histórico de queda de mais de 30 aviões.

Arquitetura Genérica de SA

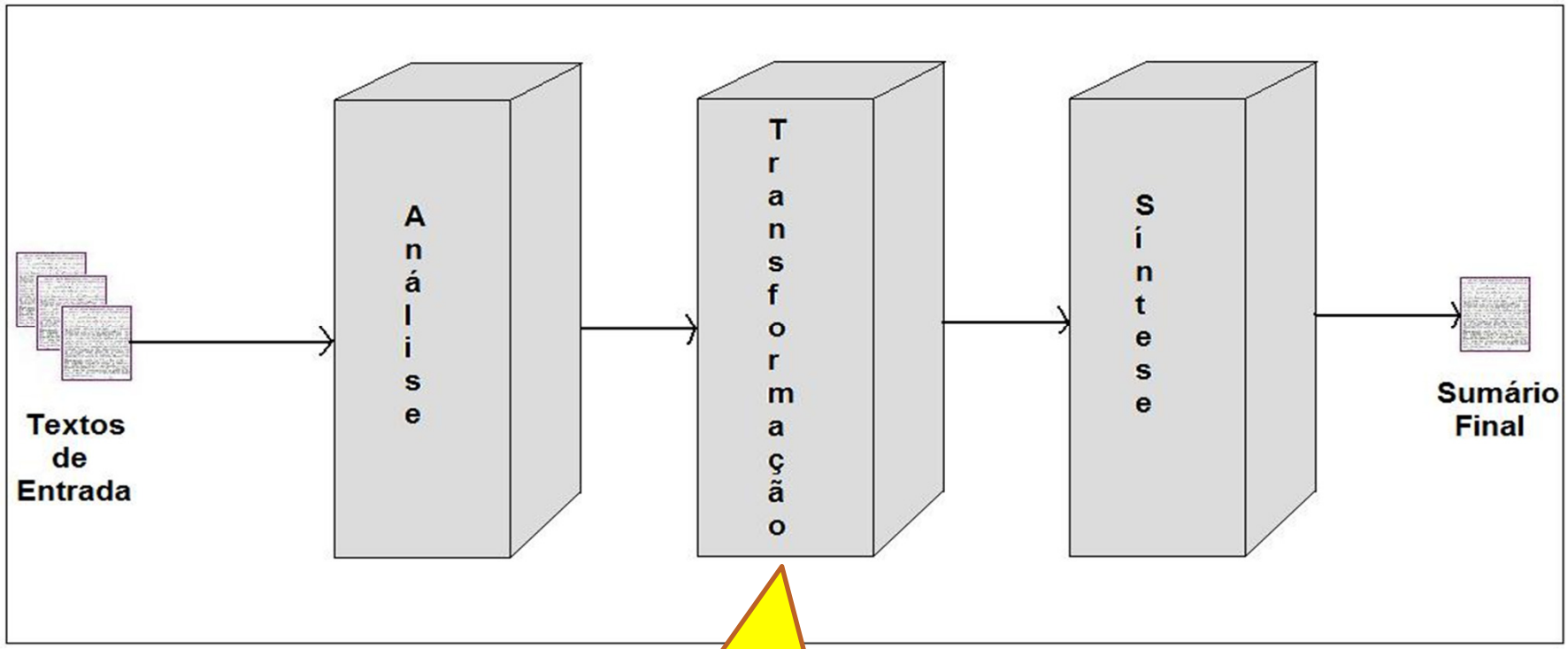


Arquitetura Genérica de SA



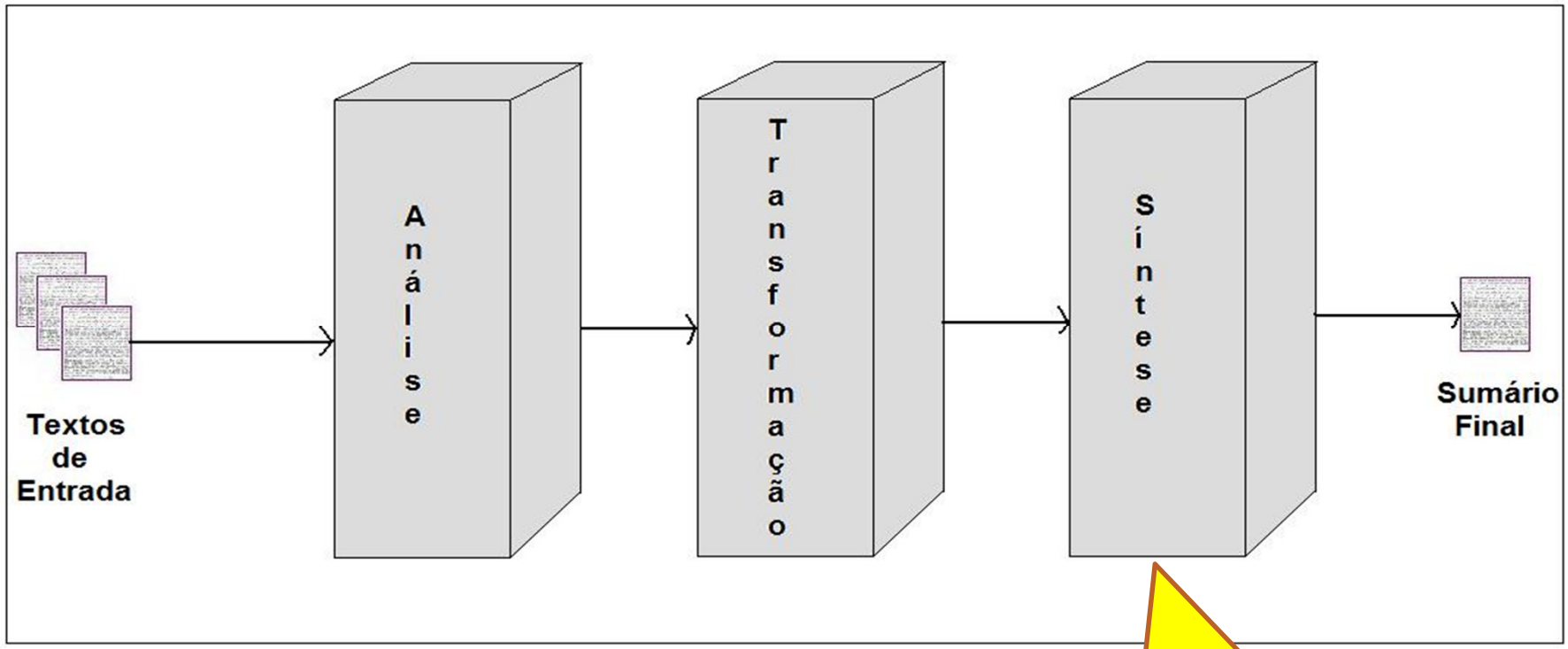
Análise CST dos textos de entrada
Segmentação e detecção topical
Resolução de expressões temporais
Resolução de correferências

Arquitetura Genérica de SA



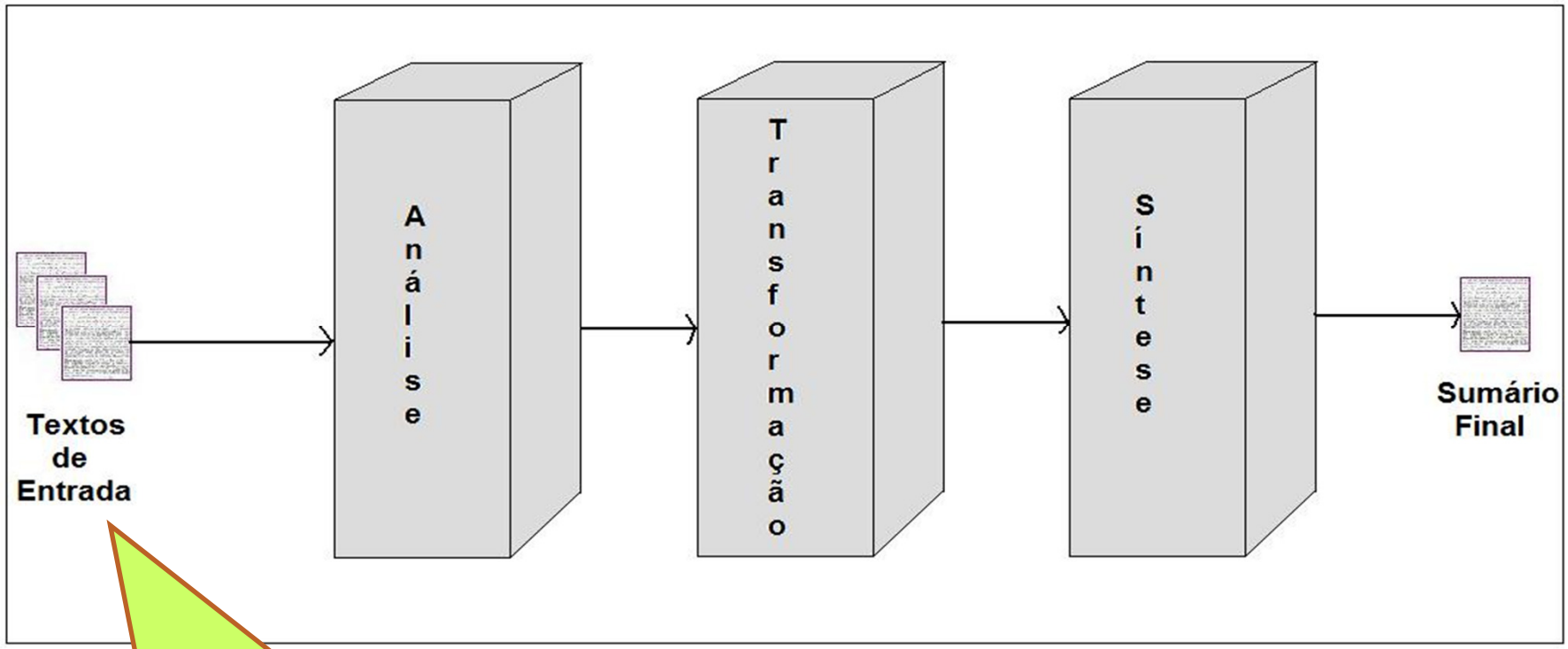
Seleção de conteúdo relevante
Ranqueamento da informação

Arquitetura Genérica de SA



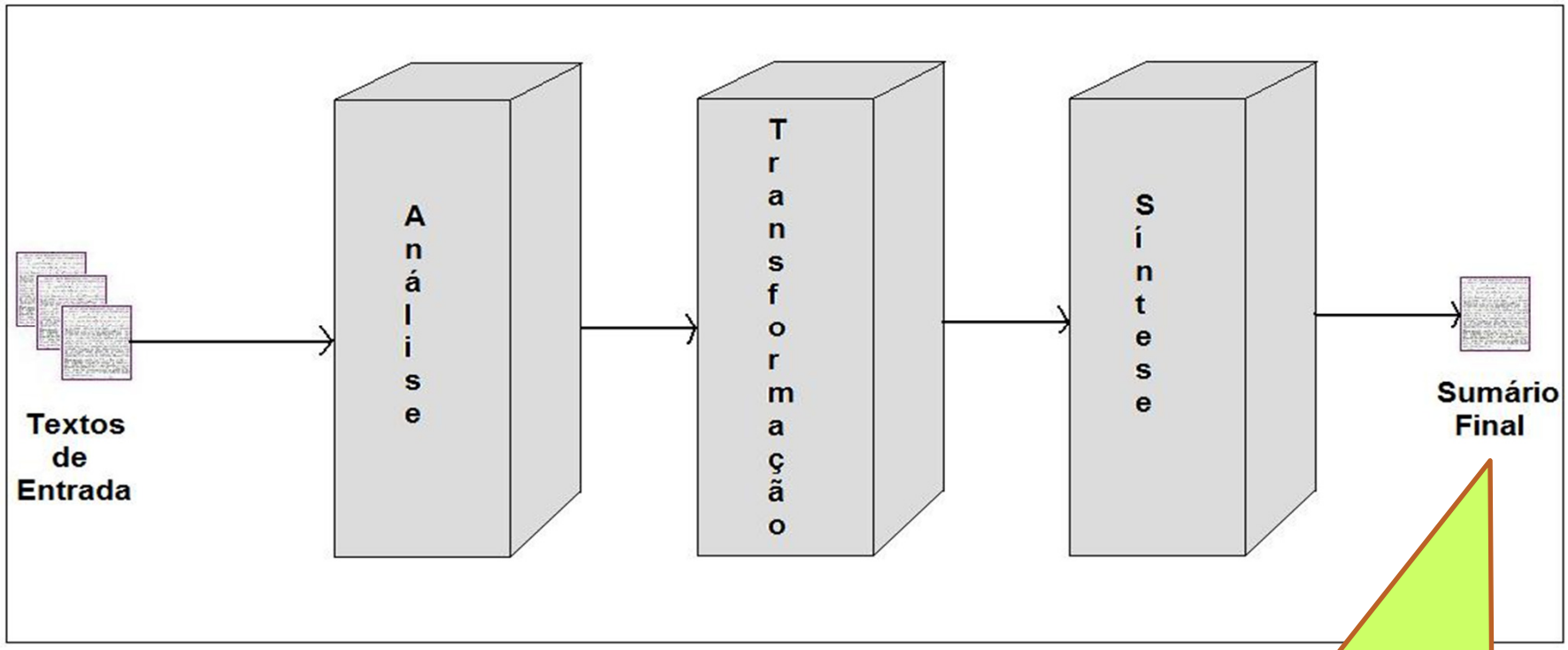
Fusão de informações
Ordenação de sentenças
Seleção de expressões referenciais

Arquitetura Genérica de SA



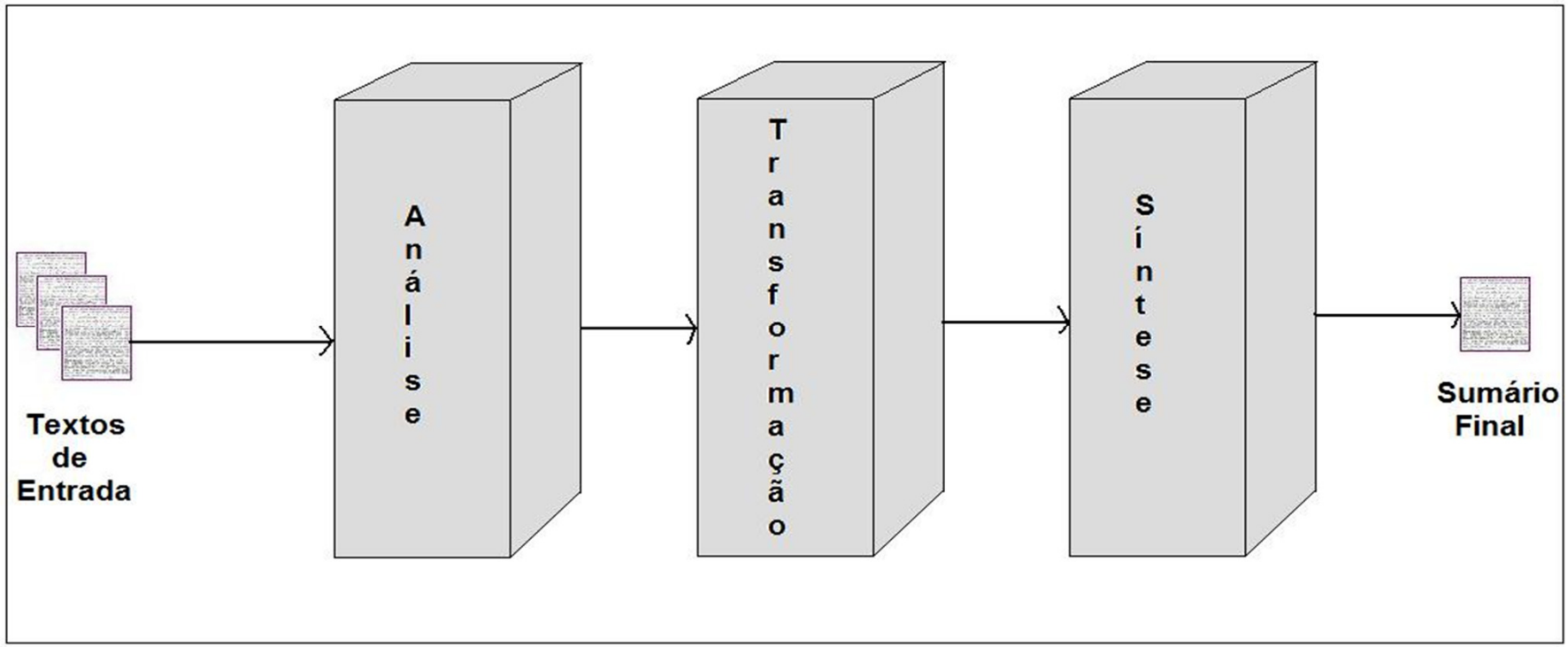
Recuperação de documentos por tópico
Agrupamento de documentos em sub-tópicos

Arquitetura Genérica de SA



Indexação de termos/sentenças/entidades
Formas de apresentação do sumário

Arquitetura Genérica de SA



Impacto da **CST** nessas etapas!

DESAFIOS MULTIDOCUMENTO VS. CST

- Como tratar os desafios usando CST?
  - Evolução de eventos no tempo
  - Narração dos eventos com diversos estilos, perspectivas diferenciadas e em momentos variados
  - Diferentes focos sobre uma mesma informação central
  - Expressões referenciais diferentes
  - 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
  - Etc.

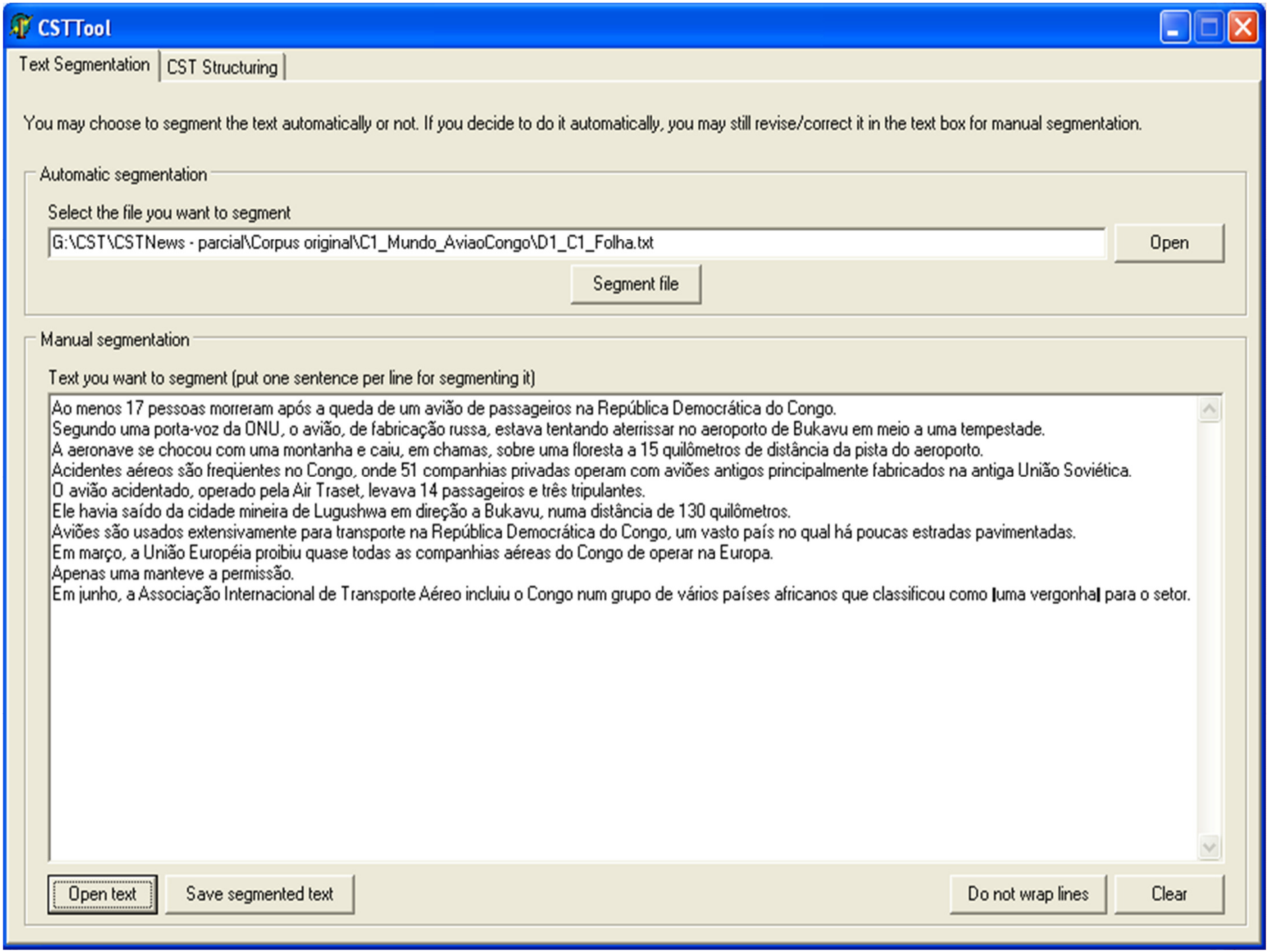
ETAPAS DO PROJETO

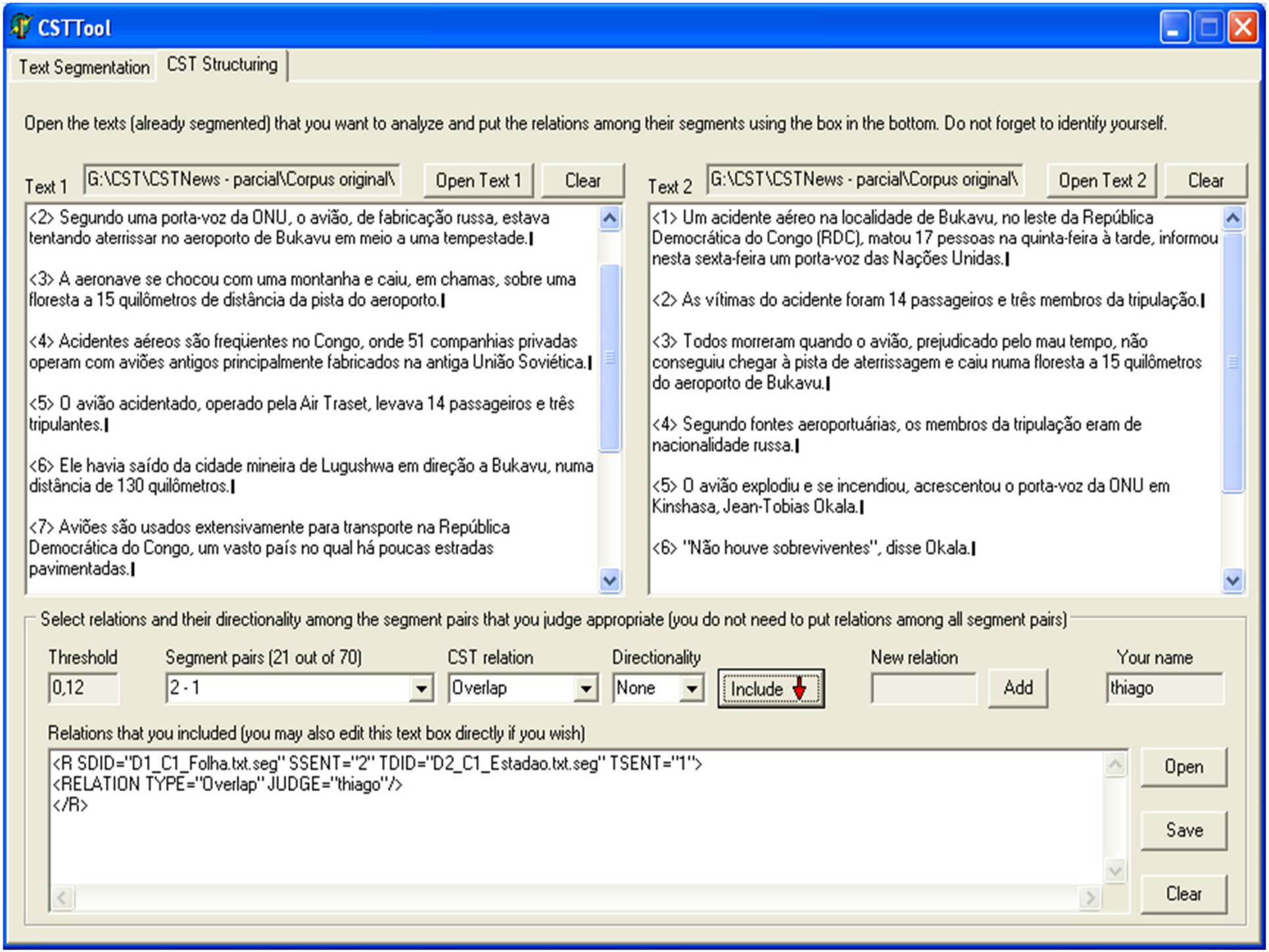
- Criação de córpus
  - Ferramentas de edição
- Análise de dados e estudos de caso
  - Testes de conceitos
- Proposta de métodos para análise, transformação e síntese

CST para o português

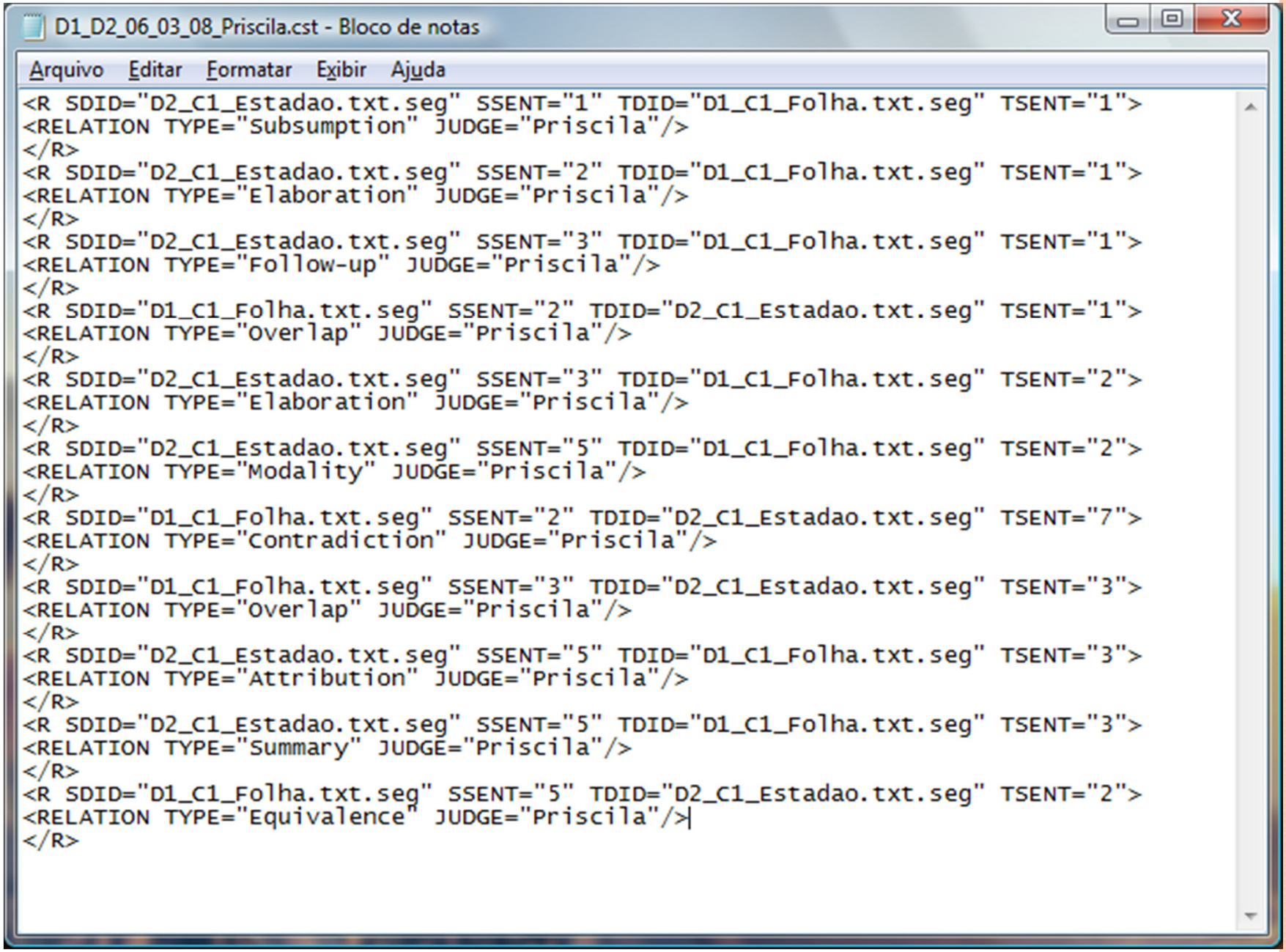
- CSTTool: uma ferramenta semi-automática para análise CST
  - Segmentação sentencial, detecção de pares de segmentos relacionados, atribuição de relações
  - XML nos moldes do CSTBank

- Futuramente, uma ferramenta completamente automática
  - *Parser multidocumento*


CSTTool
Text Segmentation
CST Structuring
You may choose to segment the text automatically or not. If you decide to do it automatically, you may still revise/correct it in the text box for manual segmentation.
Automatic segmentation
Select the file you want to segment
G:\CST\CSTNews - parcial\Corpus original\C1_Mundo_AviaoCongo\D1_C1_Folha.txt
Open
Segment file
Manual segmentation
Text you want to segment (put one sentence per line for segmenting it)
Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.
Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.
A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.
Acidentes aéreos são freqüentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.
O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes.
Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 quilômetros.
Aviões são usados extensivamente para transporte na República Democrática do Congo, um vasto país no qual há poucas estradas pavimentadas.
Em março, a União Européia proibiu quase todas as companhias aéreas do Congo de operar na Europa.
Apenas uma manteve a permissão.
Em junho, a Associação Internacional de Transporte Aéreo incluiu o Congo num grupo de vários países africanos que classificou como |uma vergonha| para o setor.
Open text
Save segmented text
Do not wrap lines
Clear


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
G:\CST\CSTNews - parcial\Corpus original\
Open Text 1
Clear
<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.|
<7> Aviões são usados extensivamente para transporte na República Democrática do Congo, um vasto país no qual há poucas estradas pavimentadas.|
Text 2
G:\CST\CSTNews - parcial\Corpus original\
Open Text 2
Clear
<1> Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.|
<2> As vítimas do acidente foram 14 passageiros e três membros da tripulação.|
<3> Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu.|
<4> Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.|
<5> O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.|
<6> "Não houve sobreviventes", disse Okala.|
Select relations and their directionality among the segment pairs that you judge appropriate (you do not need to put relations among all segment pairs)
Threshold
0,12
Segment pairs (21 out of 70)
2 - 1
CST relation
Overlap
Directionality
None
Include
New relation
Add
Your name
thiago
Relations that you included (you may also edit this text box directly if you wish)
<R SDID="D1_C1_Folha.txt.seg" SSENT="2" TDID="D2_C1_Estadao.txt.seg" TSENT="1">
<RELATION TYPE="Overlap" JUDGE="thiago"/>
</R>
Open
Save
Clear

Exemplo de anotação


D1_D2_06_03_08_Priscila.cst - Bloco de notas

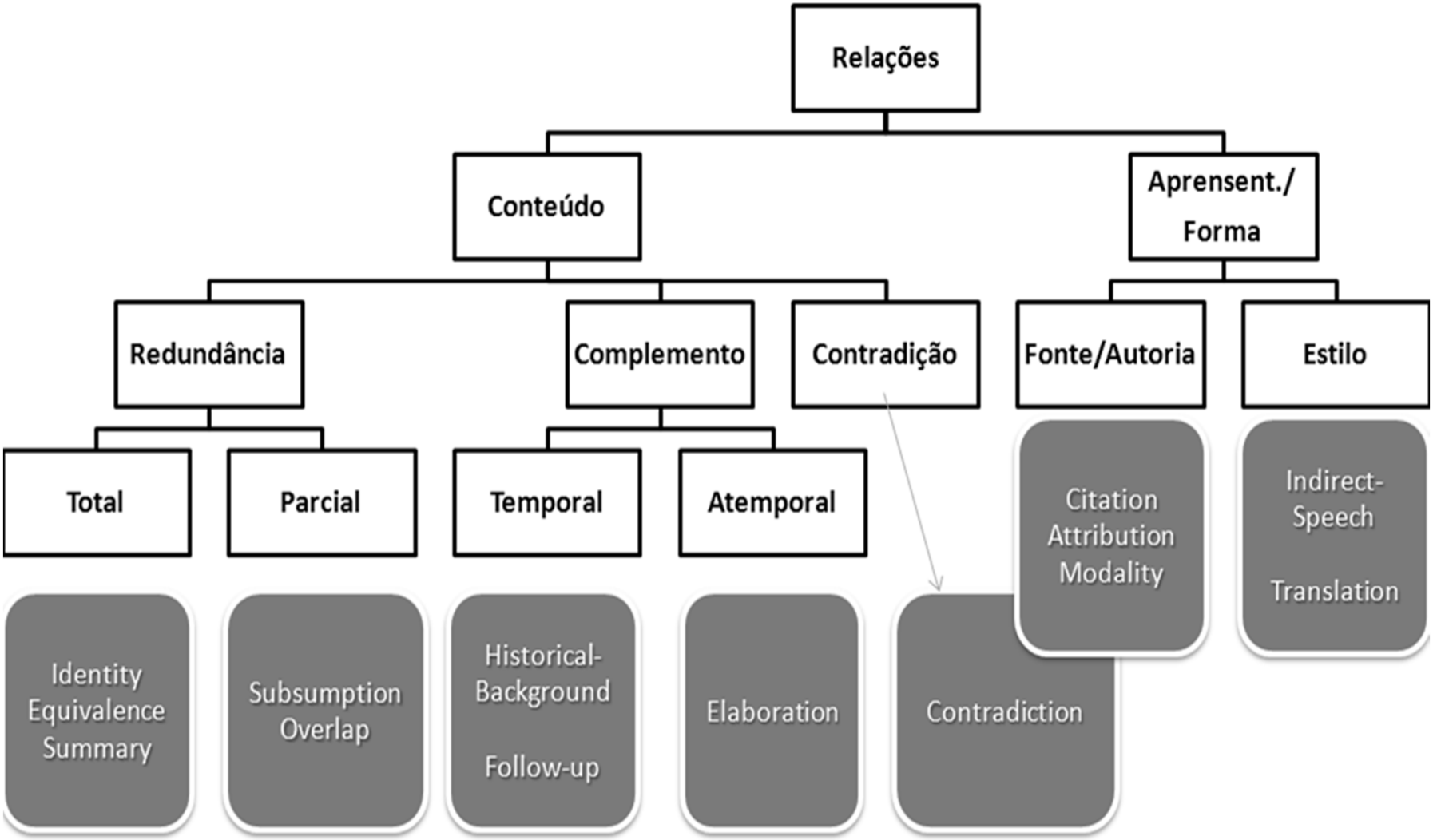
Arquivo Editar Formatar Exibir Ajuda

```
<R SDID="D2_C1_Estadao.txt.seg" SSENT="1" TDID="D1_C1_Folha.txt.seg" TSENT="1">
<RELATION TYPE="Subsumption" JUDGE="Priscila"/>
</R>
<R SDID="D2_C1_Estadao.txt.seg" SSENT="2" TDID="D1_C1_Folha.txt.seg" TSENT="1">
<RELATION TYPE="Elaboration" JUDGE="Priscila"/>
</R>
<R SDID="D2_C1_Estadao.txt.seg" SSENT="3" TDID="D1_C1_Folha.txt.seg" TSENT="1">
<RELATION TYPE="Follow-up" JUDGE="Priscila"/>
</R>
<R SDID="D1_C1_Folha.txt.seg" SSENT="2" TDID="D2_C1_Estadao.txt.seg" TSENT="1">
<RELATION TYPE="Overlap" JUDGE="Priscila"/>
</R>
<R SDID="D2_C1_Estadao.txt.seg" SSENT="3" TDID="D1_C1_Folha.txt.seg" TSENT="2">
<RELATION TYPE="Elaboration" JUDGE="Priscila"/>
</R>
<R SDID="D2_C1_Estadao.txt.seg" SSENT="5" TDID="D1_C1_Folha.txt.seg" TSENT="2">
<RELATION TYPE="Modality" JUDGE="Priscila"/>
</R>
<R SDID="D1_C1_Folha.txt.seg" SSENT="2" TDID="D2_C1_Estadao.txt.seg" TSENT="7">
<RELATION TYPE="Contradiction" JUDGE="Priscila"/>
</R>
<R SDID="D1_C1_Folha.txt.seg" SSENT="3" TDID="D2_C1_Estadao.txt.seg" TSENT="3">
<RELATION TYPE="Overlap" JUDGE="Priscila"/>
</R>
<R SDID="D2_C1_Estadao.txt.seg" SSENT="5" TDID="D1_C1_Folha.txt.seg" TSENT="3">
<RELATION TYPE="Attribution" JUDGE="Priscila"/>
</R>
<R SDID="D2_C1_Estadao.txt.seg" SSENT="5" TDID="D1_C1_Folha.txt.seg" TSENT="3">
<RELATION TYPE="Summary" JUDGE="Priscila"/>
</R>
<R SDID="D1_C1_Folha.txt.seg" SSENT="5" TDID="D2_C1_Estadao.txt.seg" TSENT="2">
<RELATION TYPE="Equivalence" JUDGE="Priscila"/>
</R>
```

CST para o português

- Anotação de um córpus de textos jornalísticos em português
  - 50 textos, 2 anotadores
  - Concordância relativamente baixa, como ocorreu no inglês
- Estudo, redefinição e formalização da CST, com posterior revisão do córpus anotado
  - Várias versões de um guia de anotação
  - Refinamento da CSTTool
  - Inúmeras reuniões de treinamento e anotação (4 anotadores)

CST: TIPOLOGIA DE RELAÇÕES



Relações CST

- Exemplo de definição

Nome da Relação: *Subsumption*
Tipo da Relação: Conteúdo->Redundância->Parcial
Direcionalidade: S1->S2
Restrições: S1 apresenta as informações contidas em S2 e informações adicionais.
Comentários: S1 contém X e Y, S2 contém X.

Relações CST e o córpus CSTNews

- Organização e definição

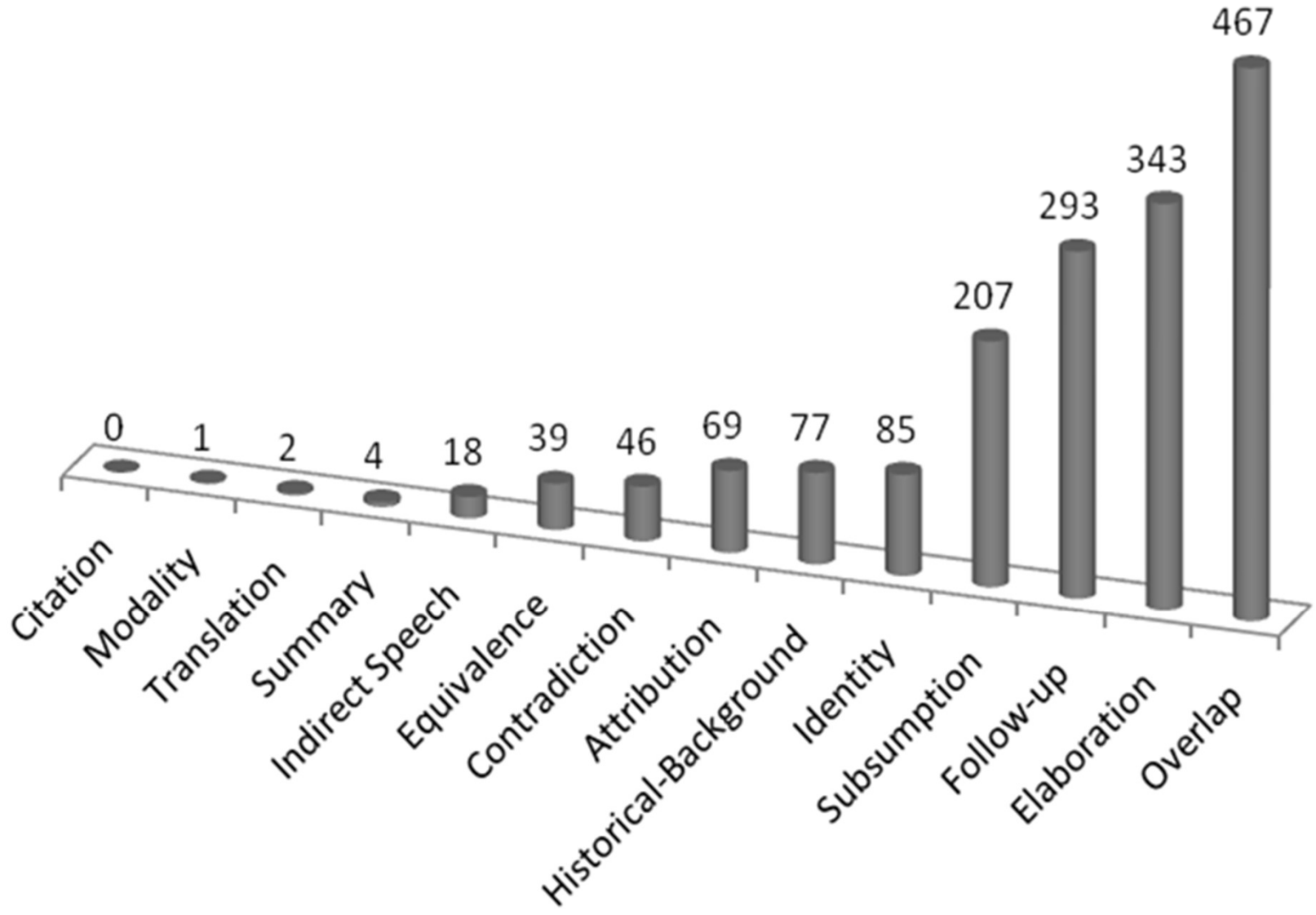
  - Relações, princípios e expressividade da teoria mantidos

	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 inglês

kappa 96% melhor do que anotação original para o porutuguês

CSTNews


0
1
2
4
18
39
46
69
77
85
207
293
343
467
Citation
Modality
Translation
Summary
Indirect Speech
Equivalence
Contradiction
Attribution
Historical-Background
Identity
Subsumption
Follow-up
Elaboration
Overlap

Parsing multidocumento

- Primeiros experimentos

  - Aprendizado de máquina
    - Instâncias: pares de segmentos
    - Classes: relações
    - Atributos superficiais
      - Diferença de tamanho em palavras
      - Sentenças idênticas
      - Posição da sentença no texto
      - Número de substantivos comuns, nomes próprios, advérbios, adjetivos, verbos, numerais
    - Testes com e sem balanceamento de classes

Parsing multidocumento

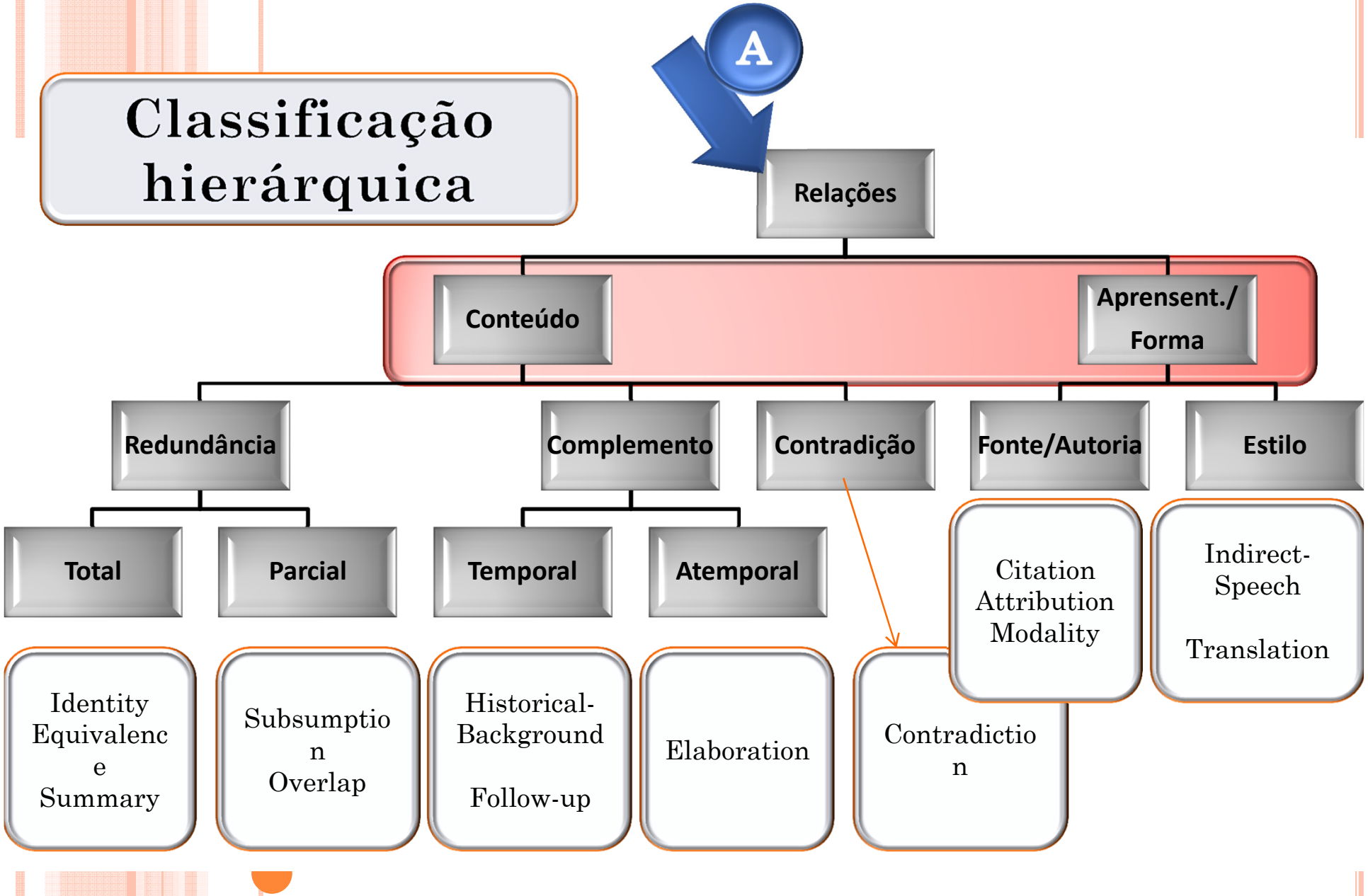
- Somente relações de conteúdo de redundância parcial e total: 77% de acerto com J48
  - *Subsumption*, *overlap*, *identity*, *equivalence*
- Todas as relações de conteúdo: 41% de acerto com J48
  - *Subsumption*, *Overlap*, *Identity*, *Equivalence*, *Elaboration*, *Follow-up*, *Historical-background*, *Contradiction*, *Summary*
- Todas as relações: 41% de acerto com J48
- Para o inglês
  - Zhang et al. (2003, 2004) com algumas relações: 29%

Parsing multidocumento

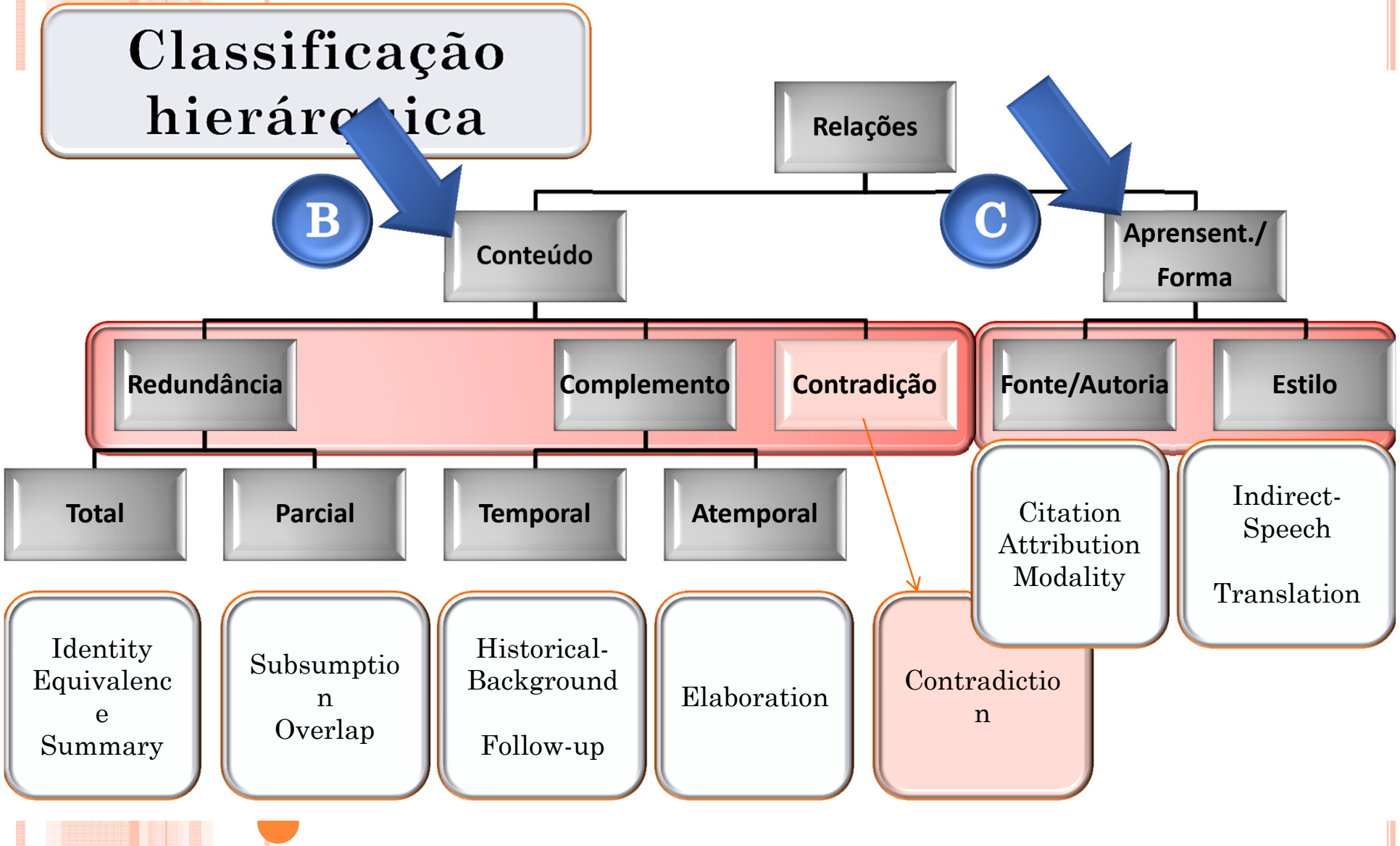
- Português vs. inglês
  - Apenas relações correspondentes

	Inglês	Português
Subsumption	0.05	0.47
Overlap	0.43	0.42
Equivalence	0.34	0.48
Elaboration	0.24	0.35
Follow-up	0.39	0.33

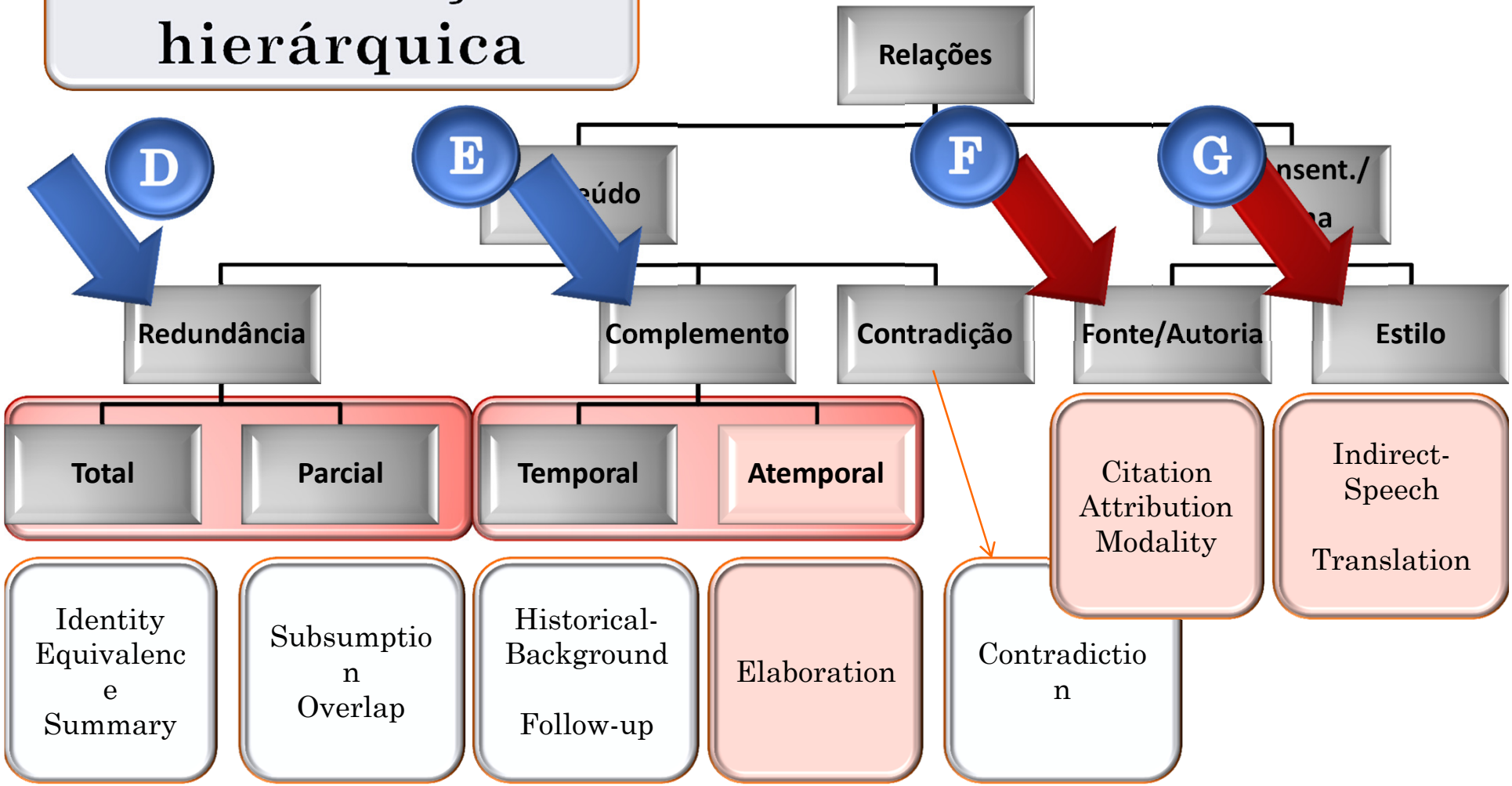
Classificação hierárquica



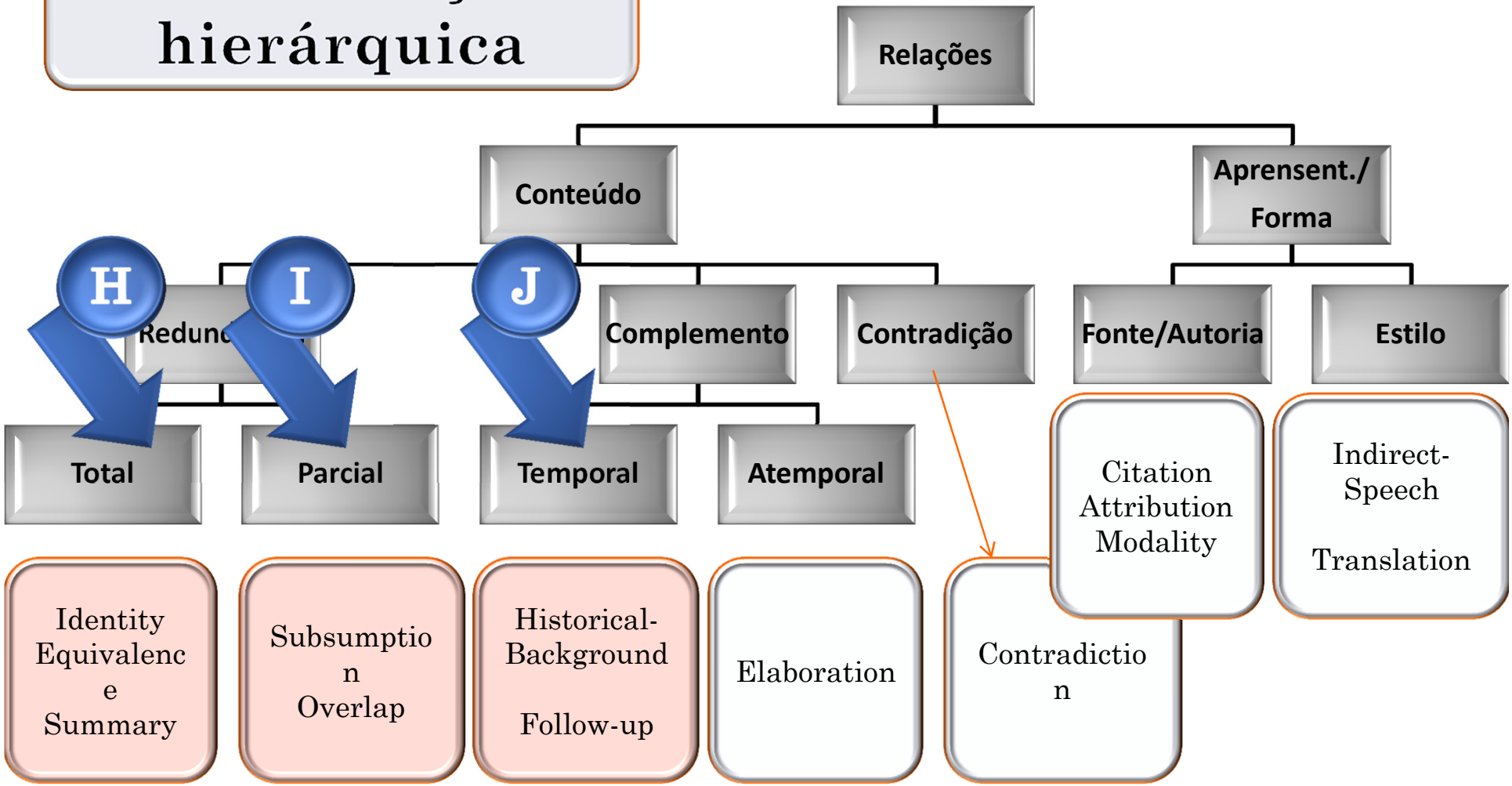
Classificação hierárquica



Classificação hierárquica



Classificação hierárquica



RESULTADOS

Classificação hierárquica

	Relação	Precisão	Cobertura	Medida-F
A	Conteúdo	0.934	0.964	0.949
	Apresentação	0.819	0.703	0.756
B	*Média*	*0.912*	*0.915*	*0.913*
	Redundância	0.697	0.726	0.711
C	Complemento	0.676	0.645	0.66
	Média	*0.687*	*0.688*	*0.687*
	Fonte	0.838	0.886	0.861
D	Estilo	0.778	0.7	0.737
	Média	*0.816*	*0.818*	*0.816*
	Parcial	0.959	0.961	0.96
E	Total	0.949	0.945	0.947
	Média	*0.954*	*0.954*	*0.954*
	Atemporal	0.627	0.618	0.623
F	Temporal	0.651	0.659	0.655
	Média	*0.639*	*0.64*	*0.639*
	Attribution	0.986	1	0.993
	Modality	0	0	0
	Média	*0.972*	*0.986*	*0.979*

	Relação	Precisão	Cobertura	Medida-F
G	Indirect	0.947	1	0.973
	Translation	1	0.875	0.933
	Média	*0.964*	*0.962*	*0.961*
H	Identity	0.932	0.965	0.948
	Equivalence	0.818	0.692	0.75
	Summary	0.684	0.813	0.743
	Média	*0.872*	*0.871*	*0.869*
I	Subsumption	0.796	0.809	0.802
	Overlap	0.828	0.816	0.822
	Média	*0.813*	*0.813*	*0.813*

Parsing multidocumento

- Versão mais sofisticada em desenvolvimento
  - Segmentação e detecção topical
  - Resolução temporal
  - Manufatura de regras “simbólicas”
  - Semântica

Seleção de conteúdo

- Proposta de operadores de seleção de conteúdo

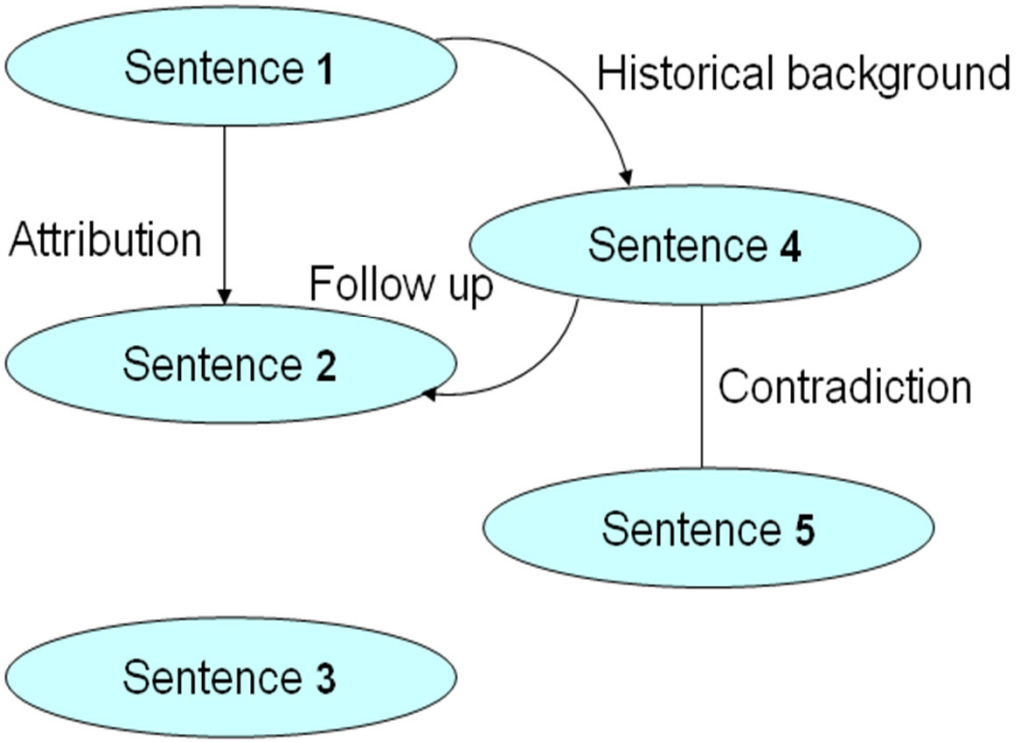
  - Estrutura CST → ranque de sentenças de acordo com informatividade

    - 6 operadores
      - Operador genérico/inicial
      - Operador de redundância
      - Operadores de contexto, contradição, autoria e desenvolvimento de eventos
        - Preferências do usuário

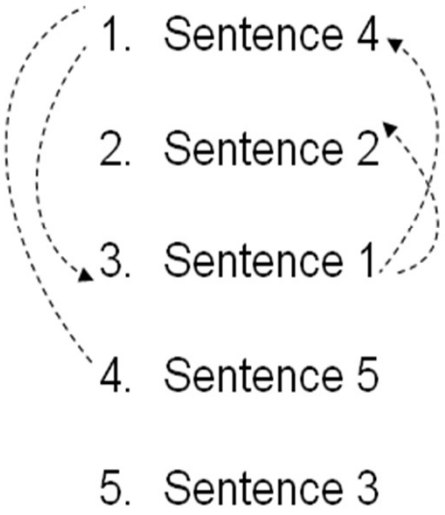
Seleção de conteúdo

- Exemplo

Estrutura CST

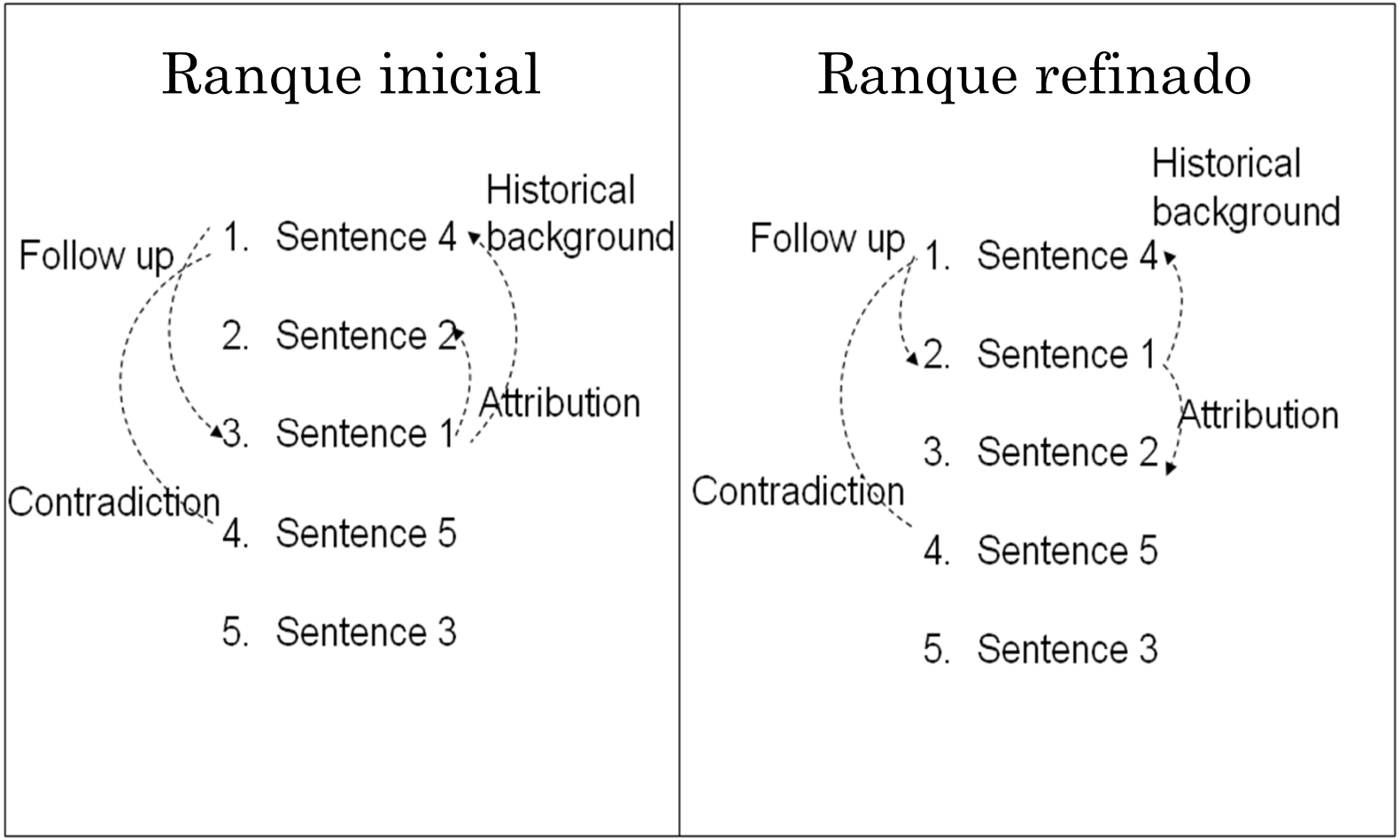


Ranque inicial



Seleção de conteúdo

- Exemplo: preferência por contexto



Avaliação dos operadores

- ROUGE (Lin e Hovy, 2003)
  - Comparação com sumários de referência (genéricos)
  - Julgamento humano da qualidade dos sumários

	Precisão	***Cobertura***	***Medida-F***
Genérico/inicial	0.5564	0.5303	0.5356
Redundância	0.5761	0.5065	0.5297
Contexto	0.5196	0.4938	0.4994
Autoria	0.5563	0.5224	0.5310
Contradição	0.5503	0.5379	0.5355
Eventos	0.5159	0.5222	0.5140

Operadores e métodos superficiais

- Adição de CST a sumarizadores superficiais
  - MEAD (Radev et al., 2000) e GistSumm (Pardo, 2005)

	Precisão	***Cobertura***	***Medida-F***
MEAD sem CST	0.5242	0.4602	0.4869
MEAD com CST	0.5599	0.4988	**0.5230**
GistSumm sem CST	0.3599	0.6643	0.4599
GistSumm com CST	0.4945	0.5089	**0.4994**

Avaliação

- Avaliação humana

0: Inaceitável
1: Ruim
2: Regular
3: Bom
4: Excelente

Preferências	Coerência	Coesão	Informatividade	Redundância
Genérico	3.6	3.2	3.6	1.8
Contexto	2.1	2.7	2.2	3.6
Autoria	3.3	2.4	3	2.8
Contradição	2.4	2.7	3.7	2.5
Eventos	2.1	2.5	3.2	2.6

Visão geral do projeto

- Caminho árduo e inédito no Brasil
- ... mas que precisa ser percorrido um dia
- Primeiros indícios de que vale a pena

EXEMPLOS

GistSumm

A seleção brasileira masculina de vôlei conseguiu, nesta sexta-feira, a sétima vitória consecutiva na Liga Mundial ao derrotar a Finlândia por 3 sets a 0 - parciais de 25/17, 25/22 e 25/21 -, em jogo realizado na cidade de Tampere, na Finlândia. A seleção brasileira masculina de vôlei, que é treinada por Bernardinho, venceu a Finlândia por 3 sets a 0, parciais de 25/17, 25/22 e 25/21, nesta sexta-feira, em Tampere (FIN), e manteve sua invencibilidade na Liga Mundial-06. O resultado de hoje deixou o Brasil perto de conquistar a única vaga do Grupo B da Liga Mundial. O Brasil arrasou a Finlândia no primeiro confronto entre as seleções, nesta sexta-feira, na cidade de Tampere, pela Liga Mundial de vôlei 2006, por 3 sets a 0, com parciais de 25/17, 25/22 e 25/21.

Sumarizador com CST (sem redundância)

O Brasil arrasou a Finlândia no primeiro confronto entre as seleções, nesta sexta-feira, na cidade de Tampere, pela Liga Mundial de vôlei 2006, por 3 sets a 0, com parciais de 25/17, 25/22 e 25/21. O resultado de hoje deixou o Brasil perto de conquistar a única vaga do Grupo B da Liga Mundial, que classifica o melhor de cada uma das quatro chaves, a Rússia (país-sede) e mais um time convidado pela Federação Internacional de Vôlei, para a fase final, de 23 a 27 de agosto, em Moscou (Rússia). Brasil e Finlândia se enfrentarão novamente neste sábado, às 12h30 (horário de Brasília), com transmissão ao vivo do canal de TV a cabo SporTV.

EXEMPLO

- Sem redundância, com contexto

A seleção brasileira masculina de vôlei conseguiu, nesta sexta-feira, a sétima vitória consecutiva na Liga Mundial ao derrotar a Finlândia por 3 sets a 0 - parciais de 25/17, 25/22 e 25/21 -, em jogo realizado na cidade de Tampere, na Finlândia.

O primeiro set permaneceu equilibrado até a metade. Na última parcial, a Finlândia novamente emparelhou o jogo com o Brasil até a metade do set, mas depois não conseguiu reagir e perdeu por 25 a 21.

A equipe brasileira masculina já conquistou cinco vezes a Liga Mundial --1993, 2001, 2003, 2004 e 2005.

SUMARIZAÇÃO AUTOMÁTICA MULTIDOCUMENTO PARA O PORTUGUÊS DO BRASIL

- www.icmc.usp.br/~taspardo
- www.nilc.icmc.usp.br/nilc/tools/CSTNews