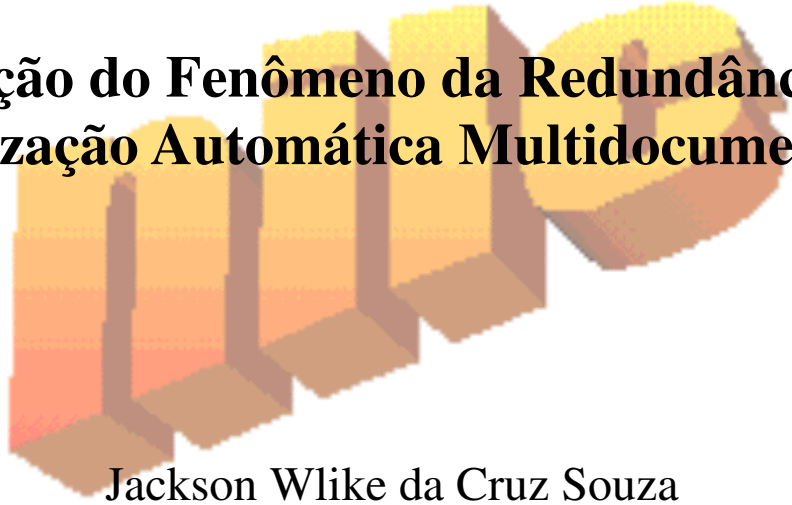
Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Investigação do Fenômeno da Redundância na Sumarização Automática Multidocumento



Jackson Wlike da Cruz Souza
Ariani Di Felippo
Thiago Alexandre Salgueiro Pardo

NILC-TR-12-03

Outubro, 2012

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Na Sumarização Automática Multidocumento (SAM), busca-se produzir automaticamente um sumário a partir de uma coleção de textos que tratam de um mesmo assunto. Especificamente, visa-se à produção de um resumo que, ao apresentar a ideia principal da coleção, seja coeso e coerente. Para tanto, a identificação e subsequente eliminação de informações redundantes ou similares advindas dos diferentes textos é fundamental. Diante da escassez de trabalhos sobre SAM que envolvem o português do Brasil (PB), foram pesquisados os principais métodos superficiais (estatísticos e linguísticos) de detecção da redundância com o objetivo de identificar os mais adequados e, assim, contribuir diretamente para o avanço das pesquisas sobre SAM em PB. Além disso, investigou-se a correspondência desses métodos com outro mais profundo, que identifica a redundância com base no tipo de relação semântica do modelo CST (do inglês, *Cross-document Structure Theory*) (RADEV, 2000). A pesquisa ora descrita foi realizada em uma iniciação científica que compreendeu o período de 01/08/2011 a 31/07/2012, sob orientação da Profa. Dra. Ariani Di Felippo e coorientação do Prof. Dr. Thiago Pardo.

O trabalho relatado contou com financiamento da FAPESP (Proc. 2011/07637-9).

Sumário

1. Introdução

Na subárea do Processamento Automático das Línguas Naturais (PLN) denominada Sumarização Automática Multidocumento (SAM), produz-se um sumário a partir de uma coleção de textos-fonte que abordam um mesmo tópico (MCKEOWN, RADEV, 1995). Nessa aplicação, o tratamento da redundância é um dos principais tópicos de investigação, pois um sumário multidocumento deve conter o conjunto de sentenças que melhor representa o tópico ou assunto da coleção sem que haja informação repetida entre elas (NEWMAN *et al.,* 2004; HENDRICKX *et al*., 2009). Na literatura, há vários trabalhos que descrevem diferentes formas, superficiais e profundas, de se detectar a redundância na SAM (p.ex.: HATZIVASSILOGLOU *et al.*, 1999, 2001; NEWMAN *et al*, 2004; HENDRICKX *et al*., 2009; JURAFSKY, MARTIN, 2009).

Diante da escassez de trabalhos sobre SAM que envolvem o português do Brasil (PB), foram pesquisados os principais métodos superficiais (estatísticos e linguísticos) de detecção da redundância entre sentenças de textos jornalísticos em PB com o objetivo de se contribuir diretamente para o avanço das pesquisas sobre SAM em PB. Além disso, investigou-se a correspondência desses métodos com outro mais profundo, que identifica a redundância com base no tipo de relação semântica do modelo CST (do inglês, *Cross-document Structure Theory*) (RADEV, 2000).

Neste relatório, descrevem-se as atividades realizadas durante todo o período de desenvolvimento desta pesquisa (01/08/2011 a 31/07/2012). Para tanto, este relatório está organizado nas seguintes Seções. Na seção 2, apresentam-se as atividades e cronograma originalmente programados. Na seção 3, descrevem-se as tarefas realizadas ao longo do projeto. Na seção 4, apresentam-se as publicações e participações em eventos resultantes do projeto ora relatado. Na seção 5, por fim, são apresentadas as considerações finais deste trabalho e os apontamentos para a continuidade da vida acadêmica do aluno.

2. Atividades e Cronograma

As tarefas previstas para a realização deste projeto de iniciação científica foram:

- **Tarefa 1:** Revisão da literatura
  Consiste no estudo da bibliografia sobre sumarização automática, redundância, métodos superficiais estatísticos e linguísticos de detecção de redundância e métodos profundos baseados na teoria/modelo CST.
- **Tarefa 2:** Seleção do *corpus*
  Consiste em selecionar o *corpus* mais adequado para servir de base à investigação em questão. No caso, o *corpus* deve apresentar as características: (i) monolíngue (PB); (ii) jornalístico, (iii) multidocumento e (iv) alinhado no nível retórico, via CST. Eventualmente, pode ser necessário complementar, adequar ou construir o *corpus* para o projeto.
- **Tarefa 3:** Teste/aplicação dos métodos de identificação da redundância
  Consiste na aplicação dos principais métodos superficiais (estatísticos e linguísticos) identificados na literatura ao *corpus* selecionado na Tarefa 2.
- **Tarefa 4:** Estudo da correlação entre os métodos superficiais e o nível de redundância
  Consiste em verificar a correlação entre os métodos aplicados na Tarefa 3 e o nível de redundância indicado pelas relações CST que une as sentenças de cada par. Em outras palavras, identificar-se-ão os métodos que expressam as diferenças de redundância.
- **Tarefa 5:** Estudo da correlação entre os métodos superficiais e as relações CST
  Consiste em verificar a correlação entre os métodos aplicados na Tarefa 3 e as relações CST anotadas entre as sentenças selecionadas a partir do *corpus*. Em outras palavras, buscar-se-á verificar quais métodos expressam adequadamente as relações CST.
- **Tarefa 6:** Escrita de relatórios e artigos científicos

3. Atividades realizadas

Nesta subseção serão detalhadas as atividades realizadas no período que engloba todo o período de pesquisa desse projeto de iniciação científica. As atividades aqui descritas compreendem as Tarefas de 1 a 6, dando ênfase aos resultados obtidos no segundo semestre da pesquisa, no qual a caracterização linguística serviu para os resultados que serão apresentados.

3.1. Revisão da literatura

Os tópicos abordados na revisão da literatura foram: (i) conceitos básicos sobre sumarização automática, (ii) fenômeno multidocumento da "redundância" e (iii) métodos de detecção de redundância superficiais e profundos (baseados na teoria CST).

3.1.1. A sumarização automática: conceitos básicos

A sumarização é uma atividade bastante comum. Na modalidade escrita, tem-se, por exemplo, notícias de jornal e as sinopses de filmes. Os sumários produzidos a partir de textos são úteis porque podem ser indexadores, permitindo que o leitor descubra o assunto do texto-fonte correspondente, ou podem ser suficientemente informativos a ponto de permitirem que o leitor dispense a leitura do texto de origem (MARTINS *et al.,* 2001). Os sumários também são úteis em várias tarefas de PLN: (i) recuperação de informação, (ii) categorização de textos, etc. Diante da utilidade dos sumários, do crescimento de informação disponível (principalmente, via *web*) e dos avanços na área de PLN, é de grande interesse a automação do processo de sumarização, foco da subárea do PLN denominada Sumarização Automática (SA). Nela, busca-se produzir automaticamente sumários a partir de um ou mais textos-fonte, sendo um "sumário" entendido como a versão mais curta de um texto ou mais textos.

De acordo com a função, os sumários podem ser informativos, indicativos ou críticos (MANI, MAYBURY, 1999). Os informativos contêm as informações principais de um texto-fonte de forma coerente e coesa ao ponto de dispensar a leitura do texto-fonte. Os indicativos apenas dizem do que o texto-fonte trata. Os críticos apresentam a informação principal do texto-fonte e avaliações sobre ele.

Quanto à forma, os sistemas de SA podem produzir extratos ou *abstracts* (SPARCK JONES, 1993). Os extratos são sumários compostos por trechos inalterados do(s) texto(s)-fonte. Os *abstracts* apresentam partes reescritas do(s) texto(s)-fonte. Quanto ao número de textos-fonte, a SA pode ser monodocumento, a partir de um único texto, e multidocumento, a partir de uma coleção de textos (MCKEOWN, RADEV, 1995).

Quanto à abordagem, Mani (2001) destaca que a SA pode ser superficial ou profunda em função da quantidade e do nível de conhecimento linguístico envolvidos na sumarização. Na superficial, utiliza-se pouco ou nenhum conhecimento linguístico para produzir sumários; o conhecimento utilizado é o empírico/estatístico. Sumarizadores superficiais costumam produzir extratos. A abordagem profunda caracteriza-se pela utilização de conhecimento linguístico morfológico, sintático, semântico e/ou pragmático-discursivo na SA. Assim, os sumarizadores profundos podem gerar extratos e *abstracts*. As abordagens superficiais e profundas podem ser mescladas, originado abordagens híbridas.

Idealmente, a SA é realizada em três etapas: (i) análise dos textos-fonte, em que se produz uma representação completa de seu conteúdo; (ii) transformação, em que o conteúdo completo do texto-fonte é condensado e (iii) síntese, em que o conteúdo condensado é expresso em língua natural na forma de um sumário (MANI, 2001). Essas etapas devem ser guiadas pela taxa de compressão, ou seja, pelo tamanho desejado do sumário. No caso da SAM, em que dois ou mais textos sobre um mesmo assunto,

provenientes de fontes distintas, são condensados em um único sumário, há uma enorme quantidade de informação redundante com a qual os sumarizados se deparam na etapa de transformação.

3.1.2. O fenômeno da "redundância"

Pode-se dizer que uma das principais diferenças entre a SA monodocumento e a SAM é o volume de informação redundante com o qual se lida durante o processo de sumarização automática. Em um conjunto de textos que tratam de um mesmo assunto, é possível encontrar um grande volume de informações em comum ou similar, assim como é possível encontrar informações que são únicas de cada texto. Consequentemente, a relação de redundância ou similaridade entre sentenças dos textos-fonte pode ser total ou parcial.

As relações de redundância total também podem ser de tipos diferentes. Dado um par de sentenças S1 e S2, as mesmas são totalmente redundantes quando: (i) S1 e S2 são idênticas, (ii) S1 e S2 apresentam o mesmo conteúdo expresso de forma diferente e (iii) S2 apresenta o mesmo conteúdo de S1 de forma compacta, com significativa diferença de tamanho. Em (i), tem-se identidade total entre forma e conteúdo. Em (ii) e (iii), tem-se identidade de conteúdo, mas não de forma.

Os exemplos do Quadro 2 ilustram esses tipos de redundância total.

Redundância total	**Sentenças**
(i) Identidade de forma e conteúdo	S1: Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.
	S2: Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.
(ii) Identidade de conteúdo	S1: Nove pessoas morreram, três delas crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão ocorrida em um mercado de Moscou, informou a polícia.
	S2: Nove pessoas morreram, sendo três crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão registrada em um mercado moscovita, informou a Polícia de Moscou.
(iii) Identidade de conteúdo	S1: De acordo com a assessoria do ministério, a transferência dos vôos de Guarulhos para Viracopos não poderá ser feita neste momento, por que o aeroporto de Campinas necessitará de ampliação, principalmente em terminal de passageiros.
	S2: Para receber os vôos de Cumbica, Viracopos precisará ser ampliado, sobretudo seu terminal de passageiros, segundo nota do Ministério da Defesa.

Quadro 2: Exemplos de redundâncias totais.

Quanto às parciais, ressalta-se que, dado um par de sentenças S1 e S2, estas apresentam redundância parciais quando: (i) S1 e S2 apresentam informações em comum e ambas apresentam informações adicionais distintas entre si (S1 contém X e Y, S2 contém X e Z) ou (ii) S1 apresenta as informações contidas em S2 e informações adicionais (S1 contém X e Y, S2 contém X). No caso (i), há uma sobreposição de informações entre S1 e S2. No caso (ii), S1 subsume S2. Os exemplos do Quadro 3 ilustram esses tipos de redundância parcial, sendo que os trechos sublinhados são redundantes e os em negrito são específicos de cada sentença.

Redundância parcial	Sentenças
(i) S1 contém X e Y, S2 contém X e Z	S1: A falha no reversor **– *mecanismo que ajuda o avião a frear* –** foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado.
	S2: O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, e ainda assim a aeronave da TAM, ***um Airbus A320***, continuou voando, com o reverso direito desligado.
(ii) S1 contém X e Y, S2 contém X.	S1: 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.**
	S2: Ao menos 17 pessoas morreram após um acidente aéreo na República Democrática do Congo.

Quadro 3: Exemplos de redundâncias parciais (MAZIERO *et al.*, 2010).

Para que os sumários multidocumentos são apresentem informações repetidas, é preciso identificar e tratar a redundância durante a etapa de transformação. Quanto à detecção, há vários métodos encontrados na literatura. A seguir, alguns deles são brevemente descritos.

3.1.3. Os métodos de detecção da redundância

Considerando-se a necessidade do tratamento da redundância, a etapa de transformação da maioria dos métodos de SAM é realizada por meio dos seguintes passos básicos (ou algoritmo), dada uma coleção de textos-fonte (RADEV *et al*., 2004): (i) calcular a importância de cada sentença dos textos-fonte; (ii) ranquear as sentenças em função de sua importância; (iii) selecionar a sentença de maior pontuação de importância para iniciar o sumário; (iv) selecionar a próxima sentença do ranque; (v) calcular a redundância ou similaridade da nova sentença candidata em relação à sentença já selecionada para o sumário; (vi) selecionar a sentença candidata para compor o sumario se esta contiver pouca sobreposição com a sentença inicialmente selecionada, e (vii) repetir os passos para as demais sentenças do ranque até que o tamanho desejado do sumário seja alcançado.

Quanto ao algoritmo de SAM, duas observações são relevantes. A primeira é referente à estratégia de ordenação das sentenças por importância. A forma de se calcular a importância de uma sentença é variada, podendo ser baseada em: (i) número de palavras-chave, (ii) frequência das palavras, (iii) localização no textos-fonte, etc. (cf. MANI, 2001). A segunda observação diz respeito à redundância. Observa-se que o algoritmo descrito engloba um "fator de redundância", que se baseia na identificação da similaridade entre uma sentença candidata e cada uma das demais sentenças já selecionadas para compor o sumário. A sentença candidata é penalizada se for muito similar às do sumário (JURAFSKY, MARTIN, 2009). Assim, a detecção da redundância é sempre feita entre duas sentenças semanticamente relacionadas, ou seja, a verificação é feita em pares. Quanto à detecção da redundância, ressalta-se que há vários trabalhos que descrevem diferentes métodos, superficiais e profundos, na SAM.

Nos trabalhos de Hatzivassiloglou *et al.* (1999, 2001), para o inglês, um método superficial estático e alguns métodos superficiais linguísticos foram analisados. O estatístico é dito tradicional e se baseia no número de palavras (de classe aberta) em comum entre as unidades de significado. A sobreposição pode ser verificada em função das formas analisadas (canônicas) ou não-analisadas (formas que ocorrem na superfície textual). Para calcular a sobreposição lexical, Hatzivassiloglou *et al.* (1999, 2001) utilizam a medida *word overlap.* O cálculo *word overlap* entre sentenças é feito por meio da aplicação da fórmula em (1).

(1)

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

Em (1), vê-se que, para calcular a *word overlap* (*Wol*) entre um par de sentenças (S1 e S2) (provenientes de textos distintos, porém que tratam do mesmo assunto), deve-se dividir o número total de palavras idênticas entre as sentenças (*CommonWords*) pela soma do número total de palavras de cada sentença (*Words(S1) + Words(S2)*), excluindo-se as *stopwords*[1], números e símbolos). O resultado obtido será entre 0 e 1, sendo que, quanto mais próximo de 1 for a *Wol,* mais redundante será o par entre si, e, quanto mais próximo de 0, menos redundante.

Hatzivassiloglou *et al.* (1999, 2001) também utilizam métodos superficiais linguísticos, os quais buscam capturar a similaridade de forma mais "inteligente". Entretanto, apesar de se basear em conhecimento linguístico mais sofisticados do que a simples sobreposição de formas lexicais, tais métodos ainda são considerados "superficiais", pois as pistas linguísticas são simples. Esses métodos, segundo os autores, são classificados em simples e compostos. Os métodos simples capturam apenas um tipo de característica das sentenças, a saber:

(i) sobreposição de etiquetas morfossintáticas: identifica etiquetas morfossintáticas em comum. Esse método é mais sofisticado que a simples sobreposição de formas linguísticas (*word overlap*). Diz-se que a medida *word overlap* é especificada para cada categoria sintática, tendo-se: *noun overlap, verb overlap, adjective overlap* e *adverb overlap.*

(ii) sobreposição de radicais (*stem*): identifica palavras que pertençam ao mesmo paradigma derivacional, ou seja, a similaridade é medida em função da sobreposição de palavras morfologicamente relacionadas. Assim, o par S1 ("O intérprete cantou de forma espetacular.") e S2 ("O cantor fez uma apresentação excelente.") é mais similar que o par S1 e S3 ("O vocalista teve um desempenho de impressionar."), já que S1 e S2 compartilham 1 caso de palavra de mesmo radical ("cantou" e "cantor" > radical *cant*), e S1 e S3, nenhum. Nesse caso, diz-se que medida em questão é a s*tem overlap.*

(iii) sobreposição de núcleos de sintagmas nominais: captura a similaridade em função de uma característica sintática das sentenças. Calcula-se a similaridade por meio da ocorrência de palavras idênticas em uma mesma posição ou função sintática, núcleo de sintagmas nominais (SN). Nesse caso, tem-se a medida *noun phrase head overlap.*

(iv) sobreposição de palavras sinônimas: busca-se identificar a similaridade em função da sobreposição de palavras semanticamente relacionadas, o caso, sinônimas. Tendo em vista esse critério, o par S1 ("O intérprete cantou de forma ***espetacular***.") e S2 ("O cantor fez uma apresentação ***excelente***.") é mais similar que o par S1 e S3 ("O vocalista teve um desempenho de impressionar."), já que S1 e S2 compartilham 2 casos de sinonímia ("intérprete"/"cantor" e "espetacular"/"excelente") e S1 e S3 apenas 1 ("intérprete"/"vocalista")[2]. Tendo em vista que a identificação da sobreposição de palavras sinônimas para o inglês é feita com base na WordNet (WN.Pr) (FELLBAUM, 1998)[3], e medida é especificada como *WordNet overlap.*

[1] As *stopwords* são basicamente palavras funcionais (p.ex.: preposições, artigos, conjunções, etc).

[2] Para esse exemplo, a sinonímia é considerada uma relação entre palavras de mesma classe gramatical, sendo que os exemplos foram elaborados com base no Tep 2 (MAZIERO *et al.,* 2008), disponível em http://www.nilc.icmc.usp.br/tep2/.

[3] A WN.Pr é uma base de dados lexicais em que as palavras e expressões do inglês norte-americano estão divididas em quatro categorias sintáticas: nome, verbo, adjetivo e advérbio. As unidades de cada categoria estão codificadas em *synsets* (do inglês, *synonym sets*), ou seja, em conjuntos de formas sinônimas ou quase-

Além dos métodos simples, Hatzivassiloglou *et al.* (1999, 2001) utilizam métodos classificados como compostos. Na verdade, tais métodos capturam dois tipos de característica das sentenças. Dentre eles, citam-se como exemplos:

(i) sobreposição de palavras + ordem: busca-se verificar se as palavras em comum em uma sentença ocorrem na mesma ordem na outra sentença do par.
(ii) sobreposição de palavras + distância entre elas: busca-se verificar se as palavras em comum ocorrem dentro de uma janela (distância) pré-definida. Caso essa janela tenha tamanho 1, focaliza-se identificar sobreposição de colocações. Caso a janela tenha tamanho 5, por exemplo, identificam-se palavras relacionadas em uma região da sentença.

Os autores ressaltam que os métodos compostos podem ser modificados considerando-se não apenas "sobreposição de palavras", mas sim a "sobreposição de etiquetas morfossintáticas" e a "sobreposição de radicais". Os autores também salientam que os métodos compostos podem ser mais sofisticados. Dado um par de sentenças, poder-se-á verificar, por exemplo, se há sobreposição de um "núcleo de SN" e de um "verbo". Essa combinação, segundo os autores, busca identificar relações gramaticais do tipo sujeito-verbo.

Além dos trabalhos de Hatzivassiloglou *et al.* (1999, 2001), Newman *et al*. (2004) também focalizam métodos de detecção da redundância no cenário da SAM. Esses autores combinam o método superficial estatístico tradicional a um método superficial linguístico, por meio do qual a similaridade é calculada com base em conhecimento de nível semântico. O método superficial linguístico, especificamente, baseia-se na identificação da sobreposição de palavras relacionadas na WN.Pr (FELLBAUM, 1998). No caso, pares de sentenças que apresentam maior número de palavras relacionadas na WN.Pr são mais similares que pares cujas sentenças apresentam menor número de palavras em comum relacionadas na base da WN.Pr (ou mesmo nenhuma sobreposição dessa natureza).

Outro trabalho a ser destacado é o de Hendrickx *et al*. (2009). Nele, os autores utilizam um método superficial linguístico na SAM de textos em holandês. Nesse método, a redundância é calculada pela similaridade semântica entre palavras alinhadas em nível sintático. Para tanto, os autores partem de um *corpus* comparável monolíngue[4] cujos textos foram manualmente alinhados no nível sentencial. Para tal alinhamento, as sentenças são submetidas a um *parser* (analisador sintático), ferramenta computacional responsável por identificar as estruturas sintáticas subjacentes às sentenças. Tais estruturas são representadas pelo *parser* em formato de árvore sintática. Na sequência, as árvores são manualmente alinhadas com o objetivo de identificar sintagmas similares. A partir do alinhamento dos sintagmas, verifica-se se as palavras que funcionam como núcleo dos sintagmas alinhados estão relacionadas na base de dados Cornetto pela sinônima e/ou pela hiponímia. A aplicação desse método parte da hipótese de que o compartilhamento de núcleos sintagmáticos semanticamente relacionados entre as sentenças de um par indica que estas são similares.

Nos métodos de SAM ditos profundos, a seleção de conteúdo é comumente feita pela identificação de relações semânticas previstas pela teoria ou modelo denominado CST (do

---

sinônimas (p.ex.: {car; auto; automobile; machine; motorcar}). Os *synsets* estão inter-relacionados pela relação léxico-semântica da antonímia e pelas relações semântico-conceituais de hiponímia, meronímia, acarretamento e causa.

[4] Um *corpus* comparável monolíngue é composto por textos originais em uma língua *x* e traduções nessa mesma língua.

inglês, *Cross-document Structure Theory*) (RADEV, 2000). A CST é um modelo semântico-discursivo multidocumento, formado por um conjunto de relações que permitem identificar similaridades, contradições, variações de estilos de escrita e informações complementares entre pares de sentenças provenientes de textos que descrevem o mesmo assunto. Para o PB, o conjunto original de relações CST foi refinado com base na anotação do *corpus* CSTNews (CARDOSO *et at.,* 2011), resultando em 14 relações (ALEIXO, PARDO, 2008): *Identity, Equivalence, Translation, Subsumption, Contradiction, Historical background, Modality, Attribution, Summary, Follow-up, Elaboration, Indirect speech, Contradiction* e *Citation*. A partir do conjunto de 14 relações, Maziero *et al.* (2010) propuseram uma tipologia. A Figura 1 ilustra a tipologia completa; relações com asterisco não têm direcionalidade.

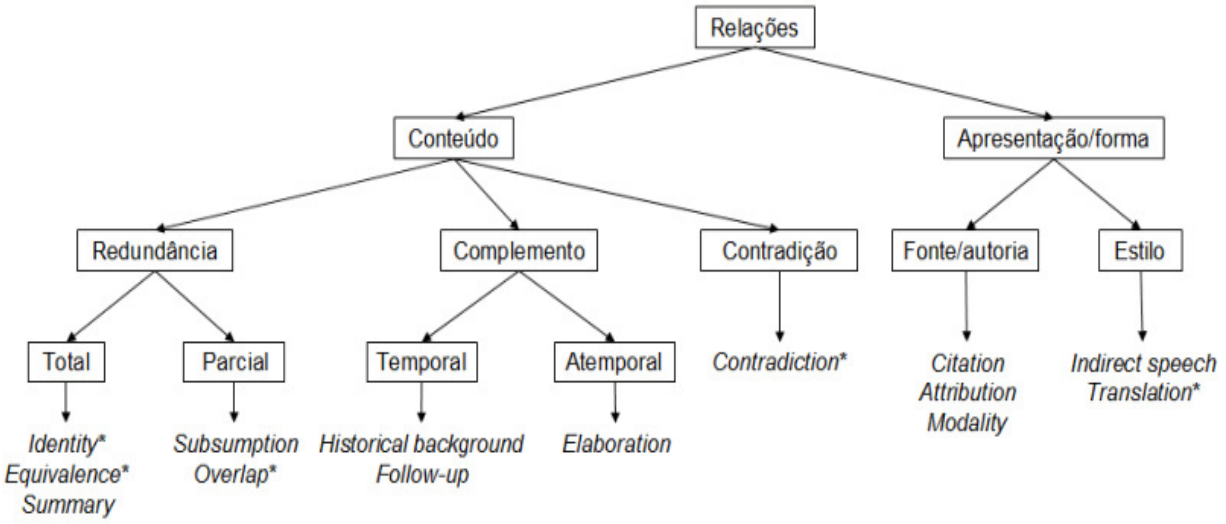


Figura 1: Tipologia das relações CST (MAZIERO *et al.,* 2010).

Essa tipologia classifica as relações em dois grandes grupos: relações de conteúdo (isto é, que ligam o conteúdo das sentenças) e relações de forma (ou seja, relações que ligam sentenças com base na forma). Cada grupo apresenta subdivisões. As relações de conteúdo podem ser classificadas nas categorias "redundância", "complemento" e "contradição". As relações da categoria "redundância", em especial, podem ser parciais ou totais, e as da categoria "complemento" podem ser temporais ou atemporais. As relações de forma, por sua vez, podem ser do tipo "fonte/autoria" ou "estilo". A SAM com base na CST pode ser ilustrada pelo algoritmo proposto pelo próprio autor da teoria, o qual prevê os seguintes passos: (i) agrupamento de textos de conteúdo similar, (ii) estabelecimento de relações CST entre as sentenças dos textos, (iii) ranqueamento das sentenças em função de algum critério de importância e (iii) seleção de sentenças para compor o sumário com base nesse ranque. Caso uma sentença tenha sido selecionada inicialmente para compor o sumário por ser a melhor colocada no ranque, verifica-se se a próxima sentença candidata está relacionada à sentença já selecionada por uma das relações de redundância. Se essa relação for de redundância total, por exemplo, a sentença candidata pode ser descartada.

A revisão dos métodos de detecção da redundância possibilitou a delimitação dos métodos investigados neste projeto, os quais estão descritos na Subseção 2.2.3 e cuja possível correspondência com as relações do tipo CST será investigada. Na sequência, descreve-se a tarefa de seleção/ construção do *corpus* que está servindo de base para a aplicação dos métodos selecionados e análise da correspondência dos mesmos com as relações CST.

3.2. Seleção do corpus

Para a execução desta pesquisa, foi necessário um *corpus.* Por definição, um *corpus* é um conjunto de dados linguísticos sistematizados de acordo com determinados critérios, de maneira que possa ser processado por computador com a finalidade de propiciar resultados vários e úteis para a descrição e análise (SINCLAIR, 2005). Por essa definição, um *corpus* é um artefato produzido para a pesquisa e que, por isso, a maioria de suas características é dependente dos objetivos da pesquisa. Tendo em vista o trabalho ora descrito, o *corpus* tinha de apresentar as seguintes características:

(i) Monolíngue: essa característica advém do fato de que se focaliza neste trabalho o estudo de métodos de SAM no cenário específico do processamento automático do PB; portanto, o *corpus* deve contemplar essa língua em questão.
(ii) Multidocumento: essa característica advém naturalmente do fato de que o foco deste trabalho está na SAM; portanto, necessitava-se de um *corpus* composto por coleções (ou *clusters*) de textos que versam sobre um mesmo assunto. Um *corpus* com essa característica permite investigar especificamente a redundância causada pelo fato de as sentenças a serem selecionadas para compor o sumário serem provenientes de mais de um texto sobre o mesmo assunto.
(iii) Alinhado/ anotação em nível retórico (via CST): essa característica advém de um dos objetivos deste trabalho, que é o de investigar a possível correlação entre os métodos superficiais de detecção da redundância e as relações CST (RADEV, 2000). Portanto, necessitava-se um *corpus* cujos textos, organizados em coleção ou *clusters,* tivessem sido relacionados por meio das relações CST.

Diante de tais características, buscou-se identificar na literatura um *corpus* que as satisfizesse. Nessa investigação, identificou-se o CSTNews (CARDOSO *et al.* 2011), único *corpus* multidocumento em PB anotado no nível semântico-retórico via CST, especificamente, pelas 14 relações estabelecidas por Aleixo e Pardo (2008). O CSTNews é composto por 50 coleções ou *clusters* de textos jornalísticos de fontes e domínios variados. Cada coleção tem em média 4 documentos coletados manualmente de jornais *online* em um período de 2 meses, em 2007. Para o fim desta pesquisa, foi feito um recorte no *corpus*, construindo-se um *subcorpus* do CSTNews. Esse recorte consistiu na seleção de pares de sentenças relacionadas pelas relações de redundância da CST. Segundo a tipologia proposta por Maziero *et al.* (2010), as relações de redundância total do modelo CST são *Identity*, *Equivalence* e *Summary* e as de redundância parcial são *Subsumption* e *Overlap*. Com base na tipologia, foi possível identificar no CSTNews pares de sentenças relacionadas especificamente por essas 5 tais relações. Além dessa seleção, foram selecionas sentenças de *clusters* distintos para a montagem de pares de sentenças completamente não-redundantes. No total, o recorte resultou em 45 pares de sentenças (Quadro 4) (Apêndice 1).

Tipo de relação	Relação	Quantidade de pares
Redundância total	*Identity*	5
	Equivalence	6
	Summary	4
Redundância parcial	*Subsumption*	8
	Overlap	8
Não-redundância	----	14

Quadro 4: Características do *subcorpus* construído para o projeto.

Tendo em vista um dos métodos superficiais de detecção da redundância investigado neste trabalho, especificamente o método linguístico "sobreposição de palavras de mesma etiqueta morfossintática", foi necessário anotar o *subcorpus* no nível morfossintático. Para tanto, optou-se por uma anotação automática, ou seja, por meio de um *tagger* (ou etiquetador), ferramenta computacional responsável por associar às palavras de um texto ou sentença uma etiqueta que indica sua correta categoria sintática no contexto. Optou-se por utilizar o etiquetador disponível no portal do LXCenter[5], o LX-Tagger (BRANCO, SILVA, 2004). Apesar de ter sido projetado para etiquetar textos em português europeu, o que pode acarretar na não etiquetação de algumas palavras características do PB, selecionou-se essa ferramenta devido à sua alta precisão (96,87%) e, sobretudo, sua interface *on-line* amigável. Quando dos raros casos de não etiquetação, as etiquetas foram inseridas manualmente. Tais casos foram identificados na fase de revisão manual da anotação. Em (2), ilustra-se a etiquetação de uma sentença do *subcorpus* pelo LX-Tagger.

(2) <s> Em_/PREP a/DA#fs capital/CAPITAL/CN#fs ,*//PNT houve/HAVER/V#ppi-3s ataques/ATAQUE/CN#mp a/DA#fs outros/OUTRO/ADJ#fs quatro/CARD#gp ônibus/ônIBU/CN#mp .*//PNT </s>

A etiquetação do LX-Tagger fornece a forma canônica das palavras (em caixa alta) e a etiqueta de categoria sintática. Além disso, o *tagger* fornece etiquetas secundárias, que indicam os traços de gênero e número, de modo-tempo e de número-pessoa. Aqui, apenas as etiquetas principais de categoria sintática foram suficientes.

Na próxima Subseção, descrevem-se os métodos superficiais de detecção da redundância que foram aplicados aos pares de sentenças do *subcorpus*.

3.3. Teste dos métodos de identificação da redundância

Com base na revisão dos principais trabalhos que focalizam o desenvolvimento de métodos de detecção da redundância (ou similaridade) no cenário da SAM, fez-se uma delimitação dos métodos investigados neste trabalho.

3.3.1. Delimitação dos métodos

Tendo em vista o caráter exploratório e inicial deste projeto, posto que não há outras pesquisas sobre esse tema no que diz respeito ao processamento do PB, e a duração e a profundidade de descrição/análise de uma iniciação científica, alguns dos métodos linguísticos delimitados são relativamente mais simples que alguns da literatura. Evitou-se selecionar métodos, por exemplo, que necessitam do alinhamento de árvores sintáticas ou da radicalização (em inglês, *stemming*) das palavras, posto tais tarefas, manuais ou automáticas, são complexas. Alem disso, optou-se apenas por métodos simples, ou seja, que capturam uma única característica das sentenças. Apesar disso, buscou-se contemplar, ainda que de forma relativamente mais simples, os principais métodos superficiais estatísticos e linguísticos aplicados na SAM em outras línguas. Vale ressaltar que tais métodos foram aplicados manualmente às sentenças do *subcorpus.*

a) Conjunto de métodos superficiais estatísticos

Esse conjunto é composto por 5 métodos: (i) *word overlap,* (ii) *noun overlap*, (iii) *verb overlap*, (iv) *adjective overlap*, e (v) *adverb overlap*.

[5] Disponível em: http://lxcenter.di.fc.ul.pt/

O método baseado na medida *word overlap* foi selecionado por ser um dos mais utilizados na literatura, sendo classificado como "clássico" para a tarefa em questão, apesar de ser reconhecidamente insuficiente (ao menos quando aplicado em isolado) para de se detectar a redundância adequada entre sentenças (HENDRICKX *et al*. 2009). O cálculo da *word overlap* foi feito com base na fórmula apresentado em (1). Os demais métodos de *word overlap,* que são especificações do método clássico em função da categoria sintática das palavras foram selecionados por serem mais sofisticados que o clássico, na medida em que buscam identificar a sobreposição de informação morfossintática entre as sentenças. Apesar de utilizar conhecimento linguístico simples, como a categoria sintática, optou-se por classificar os métodos baseados em *noun overlap*, *verb overlap*, *adjective overlap* e *adverb overlap* no conjunto de métodos estatísticos, posto que tais medidas são entendidas como variantes da *word overlap*. Para o cálculo da *word overlap* em função da categoria sintática, a fórmula em (1) foi adaptada. Em (3), ilustra-se a fórmula para o cálculo da *noun overlap* (Nol). O resultado obtido será entre 0 e 1, sendo que, quanto mais próximo de 1 for a *Wol,* mais redundante será o par entre si, e, quanto mais próximo de 0, menos redundante.

(3)

$$Nol\,(S1,S2) = \frac{\#CommonNoun}{\#Noun(S1) + \#Noun(S2)}$$

b) Conjunto de métodos superficiais linguísticos

Esse conjunto é composto por 6 métodos, a saber: (i) sobreposição de padrões morfossintáticos, (ii) sobreposição de verbo principal, (iii) sobreposição de núcleo de sujeito, (iv) sobreposição de núcleo de objeto/predicativo principal, (v) sobreposição de palavras sinônimas e (vi) sobreposição de etiquetas morfossintáticas.

O método "sobreposição de padrões morfossintáticos" busca identificar a ocorrência em comum nas sentenças de unidades lexicais complexas e colocações. O método busca capturar, de forma relativamente diferente, parte do que captura o método composto "sobreposição de palavras (de mesma etiqueta morfossintática) + distância entre elas" de Hatzivassiloglou *et al.* (1999, 2001). O método ora proposto difere do método composto desses autores por considerar apenas sequências de etiquetas com distância 1 entre elas (etiquetas em sequência direta). Em outras palavras, buscou-se identificar padrões morfossintáticos como [N_ADJ_PREP_N], [N_PREP_N_ADJ], [N_PREP_N] e [N_ADJ], etc. Para a identificação dos padrões, as *stopswods* são desconsideradas, com exceção das preposições. O cálculo da *overlap* em função dos padrões morfossintáticos (PdMol) foi feito por meio da fórmula descrita em (4). O resultado de (4) também será entre 0 e 1

(4)

$$PdMol\,(S1,S2) = \frac{\#CommonPdMorf}{\#PdMorf(S1) + \#PdMorf(S2)}$$

O método "sobreposição de verbo principal", também não citado explicitamente nos trabalhos investigados, justifica-se pelo fato de que o verbo principal em uma sentença carrega a maior carga semântica da mesma. Assim, a sobreposição do verbo principal entre duas sentenças pode indicar similaridade ou redundância entre elas. Assim, optou-se por verificar a detecção da redundância por meio desse método.

Os métodos sobreposição de "núcleo de sujeito" e "núcleo de objeto/predicativo principal" também não foram explicitamente citados na literatura. No entanto, tais métodos

são vistos como uma especificação do método "sobreposição de núcleo de SN", já que busca identificar não somente núcleos de SNs em comum, mas sim palavras que são núcleo em SNs com funções sintáticas específicas. No caso, apenas o objeto/predicativo principal foi observado.

O método "sobreposição de palavras sinônimas" foi selecionado por ser um dos mais utilizados na literatura que busca capturar a similaridade entre sentenças no nível semântico. Para a identificação da sinonímia, várias fontes de conhecimento lexical do PB digitais e impressas foram utilizadas para consulta, como: (i) o TeP 2.0, um *thesaurus* eletrônico *on-line* construído nos moldes da WN.Pr (MAZIERO *et al.,* 2008); (ii) os dicionários monolíngues *Dicionário Aurélio Eletrônico* (FERREIRA, 1999) e *Dicionário Eletrônico Houaiss da Língua Portuguesa'* (HOUAISS, VILLAR, 2001).

Ressalta-se que a delimitação dos métodos de sobreposição de "verbo principal", "núcleo de sujeito" e "núcleo de objeto/predicativo principal" também foi feita tendo em vista a concepção de similaridade de Hatzivassiloglou *et al.* (1999), segundo a qual unidades textuais (p.ex.: sentenças) são similares quando compartilham o foco em um conteúdo semântico que é expresso por participantes e ações. Assim, buscou-se garantir a detecção da redundância ou similaridade quanto a: (i) os participantes, pelos métodos "sobreposição de núcleo do sujeito" e "sobreposição de núcleo do objeto/predicativo (principal)", e (ii) a ação, pelo método "sobreposição de verbo principal".

Por fim, o método "sobreposição de etiquetas morfossintáticas", também não explicitamente citado na literatura, busca, assim como o de "sobreposição de padrões morfossintáticos", explorar a similaridade entre as sentenças com base nesse nível linguístico. Para o cálculo da *overlap* em função das etiquetas, a fórmula em (1) foi adaptada. Em (5), ilustra-se a fórmula para o cálculo dessa sobreposição (EtMol). O resultado obtido será entre 0 e 1.

(5)

$$EtMol\ (S1,S2) = \frac{\#CommonEtMorf}{\#EtMorf(S1)\ +\ \#EtMorf(S2)}$$

Aos conjuntos dos métodos estatísticos e linguísticos, acrescentou-se outro, classificado aqui como "superficial estrutural". Apesar de não ser explicitamente citado nos trabalhos revisados, o método estrutural "sobreposição de localização" fora proposto neste trabalho com base na hipótese de que a redundância entre as sentenças também pode ser capturada pela similaridade entre as posições que estas ocupam em seus textos-fonte.

Essa hipótese fora levantada com base no fato de que o conteúdo informacional em um texto está distribuído em sua estrutura de parágrafos. Os textos do tipo informativo (e gênero jornalístico), como é o caso das notícias que compõem o *subcorpus* em questão, têm a função de relatar fatos ou acontecimentos atuais, de interesse e importância para a comunidade e de fácil compreensão pelo público leitor. Para informar um fato, comumente se constrói um texto com base no método da pirâmide invertida, segundo o qual a informação é ordenada por ordem decrescente de importância. Assim, com exceção do título, uma notícia apresenta o (i) *lead*, que corresponde ao primeiro ou aos dois primeiros parágrafos do texto e expressa a informação principal a ser relatada, e (ii) o corpo do texto, que desenvolve os elementos informativos referidos no *lead* (LAGE, 2002). Consequentemete, quanto mais próximas foram as posições das sentenças em seus respectivos textos-fonte, maior a chance de serem similares.

Dessa forma, optou-se por considerar a posição de cada uma das sentenças em seu respectivo texto-fonte como um atributo para a identificação da redundância entre elas.

Essa similaridade pode ser calculada pela distância entre a posição das sentenças nos textos-fonte. Assim, quanto menor a distância entre as posições que as sentenças ocupam em seus respectivos textos-fonte, maior a redundância entre elas. O cálculo dessa distância foi assim feito: dado um par de sentenças *x* e *y,* sendo *x* a sexta sentença (S6) do Texto 1 de um *cluster* e sendo *y* a quarta (S4) do Texto 2 do mesmo *cluster,* verificou-se que a distância entre elas é igual a 2 (duas posições). Tendo em vista a variação de tamanho dos textos do *subcorpus,* a distância entre as sentenças foi normalizada, dividindo-se a distância simples entre as sentenças pela maior distância observada entre duas sentenças. O valor resultante do cálculo é entre 0 e 1, sendo que, quanto mais próximo de 1, maior a distância entre elas e menor a redundância.

(6)

$$\frac{\#Distância\ entre\ S1\ e\ S2}{\#Maior\ distância}$$

A seguir, descreve-se a caracterização linguística das sentenças necessária para a análise dos métodos de detecção da redundância aqui delimitados.

3.3.2. Caracterização linguística das sentenças dos pares

Para a aplicação dos métodos delimitados, foi necessário caracterizar cada uma das sentenças do *subcorpus* em função dos atributos relativos a esses métodos.

Para a aplicação dos métodos estatísticos, foi necessário discriminar, dada uma sentença: (i) o conjunto de palavras (de classe aberta) distintas que a constitui, (ii) o conjunto dos diferentes nomes, verbos, adjetivos e advérbios e (iii) o elenco de etiquetas morfossintáticas também distintas. Para a aplicação do método estrutural e dos linguísticos, todas as sentenças do *subcorpus* foram analisadas e descritas em função dos atributos: (i) localização no texto-fonte, (ii) padrões morfossintáticos, (iii) verbo principal, (iv) núcleo de sujeito, (v) núcleo de objeto/predicativo principal e (vi) palavras sinônimas.

Somente após essa descrição linguística foi possível aplicar os respectivos métodos de detecção da redundância, ou seja, verificar a sobreposição dos atributos entre as sentenças de cada um dos pares.

Para a caracterização linguística das sentenças, montou-se uma tabela em formato "xls", em cuja primeira coluna consta a numeração dos pares de sentenças do *subcorpus*. Nas duas colunas subsequentes, há, respectivamente, a relação CST que une as sentenças do par e as informações da origem das mesmas no CSTNews (documento e *cluster*). Na coluna seguinte, as sentenças com anotações morfossintáticas foram listadas, uma em cada linha. Os atributos foram inseridos nas demais colunas em sequência, um em cada uma delas. Nessa tabela, a explicitação da relação CST não se aplica aos pares de sentenças não-redundantes, pois estes foram extraídos de *clusters* distintos, e, por isso, não há nenhum tipo de relação entre elas.

A tabela *xls* foi dividida, para ilustração, nos Quadros 5, 6 e 7 por questão de espaço. No Quadro 5, constam os pares de sentenças anotadas, sendo que a anotação, também por questão espaço, foi simplifica por meio da exclusão das etiquetas secundárias. Especificamente, constam, no Quadro 5, um par de sentenças totalmente redundantes, um par de sentenças parcialmente redundantes e um par de sentenças não redundantes do *subcorpus*.

Par	Relação	*Cluster*	Sentenças anotadas[6]
1	*Identity* (Redundância total)	D2_C1	As/DA vítimas/*vítima*/CN de_/PREP o/DA acidente/*acidente*/CN foram/*ser*/V 14/DGT passageiros/*passageiro*/CN e/CJ três/CARD membros/*membro*/CN de_/PREP a/DA tripulação/*tripulação*/CN
		D3_C1	As/DA vítimas/*vítima*/CN de_/PREP o/DA acidente/*acidente*/CN foram/*ser*/V 14/DGT passageiros/*passageiro*/CN e/CJ três/CARD membros/*membro*/CN de_/PREP a/DA tripulação/*tripulação*/CN
17	*Subsumption* (Redundância parcial)	D3_C10	A/DA aviação/*aviação*/CN de/PREP Israel/PNM realizou/*realizar*/V durante/PREP a/DA madrugada/*madrugada*/CN de_/PREP esta/DEM segunda-feira/WD dia/*dia*/CN 7/DGT ataques/*ataques*/CN a/PREP 150/DGT alvos/*alvo*/CN em_/PREP o/DA Líbano/PNM
		D5_C10	A/DA Força/PNM Aérea/PNM israelense/*israelense*/ADJ lançou/*lançar*/V uma/UM série/*série*/CN de/PREP bombardeios/*bombardeiro*/CN contra/PREP o/DA Líbano/PNM em_/PREP esta/DEM segunda-feira/WD
35	---- (Não redundância)	D1_C31	A/DA tocha/*tocha*/CN passará/*passar*/V por/PREP vinte/CARD países/*país*/CN mas/CJ o/DA Brasil/PNM não/ADV estará/*estar*/V em_/PREP o/DA percurso/*percurso*/CN olímpico/*olímpico*/ADJ
		D2_C32	Os/DA chefes/*chefe*/CN de_/PREP a/DA usina/*usina*/CN nuclear/*nuclear*/ADJ de_/PREP o/DA Japão/PNM atingida/*atingir*/PPA por/PREP terremotos/*terremoto*/CN em_/PREP a/DA última/*último*/ADJ segunda-feira/WD admitiram/*admitir*/V que/CJ ocorreram/*ocorrer*/V mais/ADV vazamentos/*vazamento*/CN radioativos/*radioativo*/ADJ em_/PREP o/DA local/*local*/CN

Quadro 5: Exemplos de pares de sentenças do *subcorpus.*

No Quadro 6, as sentenças do Quadro 5 estão caracterizadas em função dos atributos necessários à aplicação dos métodos estatísticos, sendo que os atributos foram especificados em função da forma canônica das palavras e da exclusão das *stopwords*. Quanto ao Quadro 6, ressalta-se que o valor "0" associado a um atributo indica que uma das sentenças de um par não possui esse atributo, enquanto a outra apresenta. O símbolo "X" é associado a certos atributos quando as duas sentenças de um par não o possuem, indicando que tais atributos não se aplicam às sentenças; esse é o caso do atributo "Adjetivo" para o par 1 e do atributo "Advérbio" para os pares 1, 17 e 35 do Quadro 6.

[6] As etiquetas menos intuitivas DGT, PNM e WD, presentes no Quadro 5, expressam respectivamente: dígitos, elementos constitutivos de nomes próprios e dias da semana. É interessante notar que o conjunto de etiquetas varia entre diferentes anotadores morfossintáticos.

Par	Atributo linguístico					
	Palavra	**Nome**	**Verbo**	**Adjetivo**	**Advérbio**	**Etiqueta morfossintática**
1	vítima, acidente, ir, passageiro, membro, tripulação	vítima, acidente, passageiro, membro, tripulação	ir	X	X	DA CN PREP V DGT CJ CARD
	vítima, acidente, ir, passageiro, membro, tripulação	vítima, acidente, passageiro, membro, tripulação	ir	X	X	DA CN PREP V DGT CJ CARD
17	aviação, Israel, realizar, madrugada, segunda-feira, dia, ataque, alvo, Líbano	aviação, madrugada, dia, ataque, alvo	realizar	0	X	DA CN PREP PNM V DEM WD DGT
	força, aérea, israelense, lançar, série, bombardeio	força aérea, série, bombardeio	lançar	israelense	X	DA PNM ADJ V CN PREP DEM WD
35	tocha, passar, país, não, estar, percurso, olímpico	tocha, país, percurso	passar, estar	olímpico	X	DA CN V PREP CARD CJ PNM ADV ADJ
	chefe, usina, nuclear, Japão, atingir, terremoto, último	chefe, usina, Japão, terremoto	atingir	nuclear, último	X	DA CN PREP ADJ PNM PPA WD V CJ ADV

Quadro 6: Caracterização das sentenças: atributos necessários aos métodos estatísticos.

No Quadro 7, as sentenças do Quadro 6 estão descritas em função dos atributos necessários à aplicação do método estrutural e dos linguísticos. Para tanto, os atributos foram abreviados, a saber: localização (Loc), padrões morfossintáticos (PdMorf), verbo principal (Vp), sujeito (NSuj), objeto/predicativo principal (NObjPredp) e sinônimos (Sin). Para uma adequada interpretação dos dados do Quadro 7, ressalta-se que os padrões morfossintáticos estão associados à frequência de ocorrência dos mesmos na sentença e exemplificados por uma dessas ocorrências.

Par	Atributo linguístico					
	Loc	**PdMorf**	**NSuj**	**Vp**	**NObjPredp**	**Sin**
1	S2	CN PREP CN (2) membro da tripulação	vítima	ser	passageiro, membro	0
	S2	CN PREP CN (2) membro da tripulação	vítima	ser	passageiro, membro	0
17	S1	CN PREP CN (1) ataque a alvo	viação	realizar	ataque	0
	S1	CN PREP CN (1) série de bombardeio PNM PNM (1) força aérea	força	lançar	série	0
35	S2	N ADJ (1) percurso olímpico	tocha	passar	país	0
	S1	N ADJ (2) usina nuclear	chefe	admitir	ocorrer	0

Quadro 7: Caracterização das sentenças: atributos necessários ao método estrutural e aos linguísticos.

A seguir, na Subseção 3.3.3, descreve-se o teste propriamente dito dos métodos de detecção da redundância sobre os pares de sentenças, considerando-se, para tanto, a caracterização ilustrada nos Quadros 6 e 7.

3.3.3. Aplicação dos métodos aos pares de sentenças linguisticamente caracterizadas

Com base na descrição dos atributos de cada sentença do *subcorpus,* foi possível verificar a sobreposição entre elas em cada um dos pares. Em outras palavras, cada um dos 12 métodos delimitados na literatura (1 estrutural, 5 estatísticos e 6 linguísticos) foi manualmente aplicado a cada um dos 45 pares de sentenças com o objetivo de verificar a similaridade ou redundância entre as sentenças de cada par.

Os dados relativos à aplicação ou teste dos métodos superficiais foram organizados em um arquivo no formato "xls", cujos dados estão descritos na Tabela 1. Nessa tabela, os pares de sentenças foram inseridos na primeira coluna, cada um deles em uma linha específica. Os métodos foram inseridos nas demais colunas, cada um deles em uma coluna específica. Nos casos dos pares totalmente redundantes e parcialmente redundantes, a tabela também armazena a relação CST que une as sentenças.

Na Tabela 1, os métodos foram assim abreviados, a saber: localização (MLoc), *word overlap* (MWol)*, noun overlap* (MNol)*, verb overlap* (MVol)*, adjective overlap* (MADJol)*, adverb overlap* (MADVol)*,* padrões morfossintáticos (MPdMorf), verbo principal (MVp), núcleo de sujeito (MSuj), núcleo de objeto/predicativo principal (MObjPredp), palavras sinônimas (MSin) e etiquetas morfossintáticas (MEtMorf).

Quanto ao teste ou aplicação dos métodos, observa-se que, na Tabela 1, os valores dos métodos são de tipos diferentes. No caso, eles podem ser: (i) numéricos, como para MLoc, MWol, MNol, MVol, MADJol, MADVol, MPdMorf e MEtMorf, e (ii) categóricos, como para MPdMorf, MSuj, MVp, MObjPredp e MSin.

Para os métodos estatísticos MWol, MNol, MVol, MADJol, MADVol, MPdMorf e MEtMorf, quanto mais próximos de 1 forem os resultados dos testes, mais similares são as sentenças do par. Assim, se um par de sentenças possui MWol=1, esse valor associado ao método indica que as sentenças são totalmente similares no que tange ao número de palavras em comum. Em oposição, os resultados mais próximos de 0 indicam graus menores de similaridade entre as sentenças. Somente quanto ao método estrutural MLoc, observa-se o contrário, ou seja, os valores mais baixos de MLoc indicam maior similaridade entre as sentenças de um par.

Quando categóricos, os valores indicam se há sobreposição do atributo relativo ao método no par (valor "Sim") de sentenças ou se não há sobreposição dos atributos (valor "Não").

Além dos valores numéricos e categóricos, ressalta-se que, na Tabela 1, os métodos MADJol e MADVol podem estar associados ao valor "NA", que é uma abreviatura de "Não se Aplica". Esse valor é associado a esses dois métodos quando as sentenças de um par não são compostas por adjetivos ou advérbios; nesses casos, o valor NA indica que o método não se aplica para a verificação da redundância.

Tendo em vista que os métodos MADJol e MADVol não se aplicaram a todos os pares de sentenças, decidiu-se por não incluí-los no estudo da correlação entre os métodos e o nível de redundância e no estudo da correlação entre os métodos e as relações CST. Assim, nas análises apresentadas na próxima subseção, apenas 10 métodos foram considerados, a saber: MLoc, MWol, MNol, MVol, PdMorf, MEtMorf, MSuj, MVp, MObjPredp e MSin.

Par	Rel. CST	Método											
		Estrut.	Estatístico					Linguístico					
		MLoc	MWol	MNol	MVol	MADJol	MADVol	MPdMorf	MSuj	MVp	MObjPredp	MSin	MEtMorf
01	Ident.	0	1	1	1	**NA**	**NA**	1	Sim	Sim	Sim	Não	1
02	Ident.	0	1	1	1	1	1	1	Sim	Sim	Sim	Não	1
03	Ident.	0	1	1	1	1	1	1	Sim	Sim	Sim	Não	1
04	Ident.	0	1	1	1	1	**NA**	1	Sim	Sim	Sim	Não	1
05	Ident.	0	1	1	1	1	**NA**	1	Sim	Sim	Sim	Não	1
06	Sum.	0	0,23	1	0,66	1	0,66	0	Não	Não	Não	Não	1
07	Sum.	0,18	0,8	1	1	1	**NA**	1	Sim	Sim	Sim	Não	1
08	Sum.	0,45	0,30	0,33	0	**NA**	0,5	0	Não	Não	Não	Não	0,66
09	Sum.	0,45	0,36	0,47	0.25	**NA**	0	0	Não	Não	Não	Não	0,92
10	Equi.	0,9	0,66	1	1	0	0	1	Sim	Sim	Não	Sim	0,72
11	Equi.	0,36	0,19	0,36	0	0	0	0	Não	Não	Não	Não	0,82
12	Equi.	0,18	0,45	0,61	0,5	0	0	0,4	Não	Sim	Não	Sim	0,52
13	Equi.	0,27	0,42	0,42	0,5	0,66	0	1	Não	Sim	Não	Sim	0,77
14	Equi.	0	0,68	0,46	0,6	0,5	**NA**	0	Sim	Sim	Não	Não	0,90
15	Equi.	0,09	0,36	0,57	1	**NA**	**NA**	0	Sim	Sim	Não	Não	1
16	Subs.	0	0,47	0,46	0,5	0,4	**NA**	0,5	Sim	Não	Não	Não	0,85
17	Subs.	0	0,22	0,30	0	0	**NA**	0,5	Não	Não	Não	Não	0,88
18	Subs.	0	0,33	0,5	0	**NA**	**NA**	0	Não	Não	Não	Não	0,77
19	Subs.	0,27	0,08	0,20	0,16	**NA**	0	0	Não	Não	Não	Não	0,75
20	Subs.	0	0,2	0,12	0,4	0	0	0	Não	Não	Não	Não	0,75
21	Subs.	0,18	0,5	0,30	0,66	0,8	0,5	0	Não	Não	Não	Não	0,6
22	Subs.	0,36	0,28	0,30	0,4	0	**NA**	0	Não	Não	Não	Não	0,63
23	Subs.	1	0,15	0,13	0,33	**NA**	0	1	Não	Não	Não	Sim	0,71
24	Over.	0,36	0,44	0,66	0	0	0	0,8	Não	Não	Não	Não	0,66
25	Over.	0,36	0,44	0,47	0	0,66	0	0,66	Não	Não	Não	Não	0,5
26	Over.	0,09	0,29	0,12	0	0,5	0	1	Não	Não	Sim	Não	0,28
27	Over.	0,09	0,3	0,28	0,4	**NA**	0	0,66	Não	Não	Não	Não	0,88
28	Over.	0	0,42	0,25	0,28	0	0	0,44	Não	Não	Não	Não	0,9
29	Over.	0,18	0,42	0,6	0	**NA**	**NA**	0	Sim	Não	Não	Não	0.8
30	Over.	0	0,29	0,28	0	0,33	0	0,66	Não	Não	Não	Sim	0,77
31	Over.	0	0,2	0,44	0	0,66	0	0,66	Sim	Não	Não	Não	0.82
32	Perm.	0,18	0	0	0	0	0	0,33	Não	Não	Não	Não	0,625
33	Perm.	0	0	0	0	0	0	0,33	Não	Não	Não	Não	0,84
34	Perm.	1	0,06	0	0	0	0,5	0,4	Não	Não	Não	Não	0,8
35	Perm.	0,09	0	0	0	0	0	0,66	Não	Não	Não	Não	0,84
36	Perm.	0,72	0	0	0	0	0	0	Não	Não	Não	Não	0,77
37	Perm.	0	0,16	0,09	0,4	0	0	0,5	Não	Não	Não	Não	0,7
38	Perm.	0,9	0	0	0	0	0	0	Não	Não	Não	Não	0,66
39	Perm.	0	0	0	0	0	0	0,66	Não	Não	Não	Não	0,66
40	Perm.	0,27	0	0	0	0	0	0,5	Não	Não	Não	Não	0,66
41	Perm.	0,54	0	0	0	0	0	0	Não	Não	Não	Não	0,8
42	Perm.	0,27	0,08	0	0,5	**NA**	**NA**	0,66	Não	Não	Não	Não	0,75
43	Perm.	0	0,08	0	0,28	0	0	0	Não	Não	Não	Não	0,75
44	Perm.	1	0	0	0	0	0	0	Não	Não	Não	Não	0,66
45	Perm.	0,27	0	0	0	0	0	0,66	Não	Não	Não	Não	0,36

Tabela 1: Teste/Aplicação dos métodos superficiais.

Após a aplicação dos métodos superficiais selecionados, passou-se à investigação da correlação destes com o nível de redundância e o tipo de redundância (relações CST). Para tanto, os dados contidos na Tabela 1 foram analisados de forma manual, resultando em observações preliminares a respeito das correlações em questão, e de forma automática, como complementação às observações da análise manual.

3.3.4. Investigação da correlação entre os métodos superficiais e o nível de redundância

a) Análise manual preliminar

Para a investigação da correlação entre os métodos e o nível de redundância, salienta-se que, segundo a tipologia proposta por Maziero *et al*. (2010), o *subcorpus* do CSTNews construído para o projeto é composto por 45 pares de sentenças, sendo:

(i) 15 pares de sentenças totalmente redundantes;
(ii) 16 pares de sentenças parcialmente redundantes;
(iii) 14 pares de sentenças não redundantes (advindas de coleções distintas).

A análise manual da correlação entre os métodos superficais e o nível de redundância teve início com o cálculo da média simples dos valores obtidos por cada método em função dos pares de cada nível de redundância. Na sequência, verificou-se na Tabela 1 o número de pares que obtiveram valores iguais ou superiores à média simples. No caso do MLoc, identificou-se o número de pares que obtiveram valores iguais ou inferiores que a média simples. Por exemplo, os valores obtidos por MLoc para os 15 pares totalmente redundantes foram somados (0 + 0 + 0 + 0 + 0 + 0 + 0,18 + 0,45 + 0,45 + 0,9 + 0,36 + 0,18 + 0,27 + 0 + 0,09 = 2,88) e o resultado dessa soma (2,88) foi dividido pelo número de pares desse nível, ou seja, 15. Dessa divisão, obteve-se a média simples de 0,19. A partir dessa média simples, verificou-se na Tabela 1 que, dos 15 pares da categoria totalmente redundantes, 10 deles apresentaram MLoc ≤ 0,19 (0; 0, 0; 0; 0; 0; 0,18; 0,18; 0; 0,09), o que é representando na Tabela 2 por 10/15.

Na Tabela 2, sistematizam-se os resultados da verificação aqui descrita para os 12 métodos nos 3 diferentes níveis de redundância e se destacam os métodos que obtiveram valores iguais ou superiores (no caso de MLoc, inferiores) à média simples em mais de 50% dos pares de cada nível (ou seja, em mais de 7,5 pares totalmente redundantes, em mais de 8 pares parcialmente redundantes e em mais de 7 pares não redundantes).

Método	**Nível de redundância**		
	Total	**Parcial**	**Nula**
MLoc	10/15	9/16	9/14
MWol	8/15	7/16	4/14
MNol	8/15	6/16	1/14
MVol	8/15	7/16	3/14
MPdMorf	8/15	10/16	7/14
MSuj	9/15	3/16	0/14
MVp	11/15	0/16	0/14
MObPredp	6/15	1/16	0/14
MSin	4/15	2/16	0/14
MEtMorf	10/15	11/16	8/14

Tabela 2: Métodos *vs* níveis de redundância.

(i) Os métodos MLoc, MPdMorf e MEtMorf parecem não expressar as diferenças de redundância, pois destacaram-se nos pares dos 3 níveis. Quanto ao MLoc, ressalta-se ainda que o elevado número de pares de redundância nula (categoria "permuta") com valores iguais ou inferiores à média simples se deve ao fato de que as sentenças aleatoriamente selecionadas para compor essa categoria apresentam coincidentemente posições iguais ou próximas em seus textos-fonte. Quanto aos métodos MEtMorf e MPdMorf, salienta-se que as categorias sintáticas isoladas ou em sequência (padrão morfossintático) representam generalizações linguísticas que parecem não capturar as distinções de redundância.

(ii) Os métodos MWol, MNol e MVol parecem expressar a existência ou não de redundância, pois se destacaram em muitos pares de redundância total e parcial e em poucos pares de sentenças não redundantes. A relevância do MWol pode ser justificada pelo fato de que a ocorrência de expressões linguísticas idênticas nas sentenças de um par indica sobreposição de conteúdo entre elas, mas não indica o nível dessa redundância. O métodos MNol e MVol, ao serem pautados somente nas categorias sintáticas mais relevantes para a expressão de conteúdo, também identificam sobreposição sem distinguir o nível da mesma.

(iii) Os métodos MSuj e MVp parecem expressar a redundância total, posto que se destacaram, quanto à média simples, apenas em pares totalmente redundantes. Isso se justifica pelo fato de que o sujeito e do verbo principal são elementos centrais de uma sentença, carregando as informações linguisticamente importantes para depreensão do sentido, e, por isso, a sobreposição do núcleo do sujeito e do verbo principal entre duas sentenças indica redundância de nível total.

(iv) Os métodos MSuj, MVp, MObPredp e MSin não obtiveram valores iguais ou superiores à média simples nem nenhum dos pares da categoria "redundância nula". Assim, pode-se dizer que, diante de valores baixos atribuídos a esses métodos, não há redundância entre as sentenças.

b) Análise automática

A análise automática dos dados da Tabela 1 consistiu na geração automática de regras capazes de discriminar os diferentes níveis de redundância. Especificamente, os dados resultantes da aplicação dos métodos superficiais foram submetidos ao ambiente de Aprendizado de Máquina (AM) denominado *Weka* (*Waikato Environment for Knowledge Analysis*) (FRANK *et al.*, 2011). O *Weka* é um ambiente computacional composto por algoritmos de aprendizagem advindos de diferentes abordagens da Inteligência Artificial. Os algoritmos implementados no *Weka* são capazes de analisar automaticamente os dados de entrada e aprender padrões estatisticamente relevantes, os quais são expressos por regras.

Neste trabalho, os dados da Tabela 1 foram submetidos a vários algoritmos do *Weka*. No entanto, descrevem-se aqui as regras geradas somente pelos algoritmos que obtiveram as taxas de acerto mais elevadas, a saber: os algoritmos PART e J48. Especificamente, o PART é um algoritmo que gera regras no formato de expressões lógicas (ou seja, do tipo *se, então*).

O algoritmo PART identificou o conjunto de 5 regras sequenciais descritas em (1), com 97.7% de precisão.

(1) **PART – Redundância**

Se MNol<=0.09 então **nulo** (14)
Senão se MVp=não e MEtMorf<=0.9 e MLoc<=0.27 então **parcial** (12)
Senão se MVp=sim então **total** (11)
Senão se MPdMorf<=0.33 então **total** (5/1)
Senão **parcial** (3)

Como mencionado, tais regras são sequencias e, por isso, a interpretação das mesmas também deve ser sequencial. Quanto à redundância nula, ressalta-se que o método MNol com valor menor ou igual a 0.9 foi suficiente para caracterizar esse nível de redundância em todos os 14 pares da categoria "permuta", o que refina a observação feita em (ii). Assim, o MNol ao menos distingue a redundância nula das demais (total e parcial). Na sequência, diante de MNol maior que 0.9, os métodos MVp igual a "não", MEtMorf menor ou igual a 0.9 e MLoc menor ou igual a 0.27, em conjunto, foram suficientes para caracteriza 12 dos 16 pares de redundância parcial. Já diante de MNol maior que 0.9 e de ao menos um dos métodos MVp=não e MEtMorf<=0.9 e MLoc<=0.27 com valor diferente a esses previstos, a redundância foi total diante de MVp=sim. Por fim diante de MNol maior que 0.9, ao menos um dos métodos MVp=não e MEtMorf<=0.9 e MLoc<=0.27 com valor diferente a esses previstos, MVp=sim e MPdMorf=0.33, a redundância também foi total. Segundo essa regra, no entanto, 1 par de sentenças parcial foi identificado equivocadamente como total. O par em questão é o de número 22, descrito abaixo. Caso nenhuma das regras se aplique, a redundância padrão identificada foi "parcial".

Par	Relação	Cluster	Sentenças originais
22	Subs.	D2_C15	Nove pessoas morreram, três delas crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão ocorrida em um mercado de Moscou, informou a polícia.
		D3_C15	Anteriormente, a Polícia havia informado sobre nove mortos, sendo três deles crianças, e 25 feridos.

Par	Rel. CST	Método											
		Estrut.	Estatístico					Linguístico					
		MLoc	MWol	MNol	MVol	MADJol	MADVol	MPdMorf	MSuj	MVp	MObjPredp	MSin	MEtMorf
22	Subs.	0,36	0,28	0,30	0,4	0	NA	0	Não	Não	Não	Não	0,63

O algoritmo J48 gerou o conjunto de regras descritas em (2), com 95.5% de precisão. Observa-se que as regras do J48 são semelhantes ao do PART, mas apresentam menor precisão (95.5<97.7). No entanto, o J48 pauta-se em apenas 4 métodos, enquanto o PART, em 5 métodos diferentes.

(2) **J48 – Redundância**

MNol <= 0.09 então **nulo** (14.0)
MNol > 0.09
    MVp = sim então **total** (11.0)
    MVp = não
        MEtMorf <= 0.9 então **parcial** (18.0/2.0)
        MEtMorf > 0.9 então **total** (2.0)

Esse algoritmo permite que as regras sejam visualizadas em árvore. No caso, a árvore relativa às regras do J48 para "métodos vs redundância" neste trabalho é a da Figura 2.

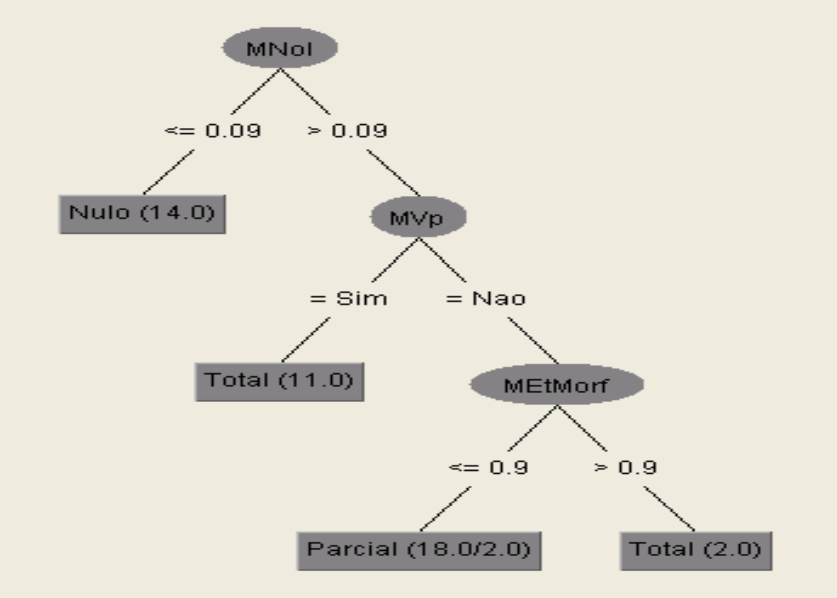


Figura 2: Visualização gráficas das regras geradas pelo J48.

3.3.5. Investigação da correlação entre os métodos superficiais e as relações CST

a) Análise manual

A análise manual da correlação entre os métodos superficais e as relações CST também teve início com o cálculo da média simples dos valores obtidos por cada método em função dos pares de cada tipo de relação. O cálculo da média simples seguiu o mesmo procedimento adotado na análise manual da correlação entre os métodos e o nível redundância. Na sequência à identificação das médias simples, verificou-se na Tabela 1 o número de pares que obtiveram valores iguais ou superiores à média simples.

Na Tabela 4, sistematizam-se os resultados da verificação aqui descrita para os 12 métodos nas 5 diferentes relações CST e nos pares de permuta. Ademais, destacam-se os métodos que obtiveram valores iguais ou superiores à média simples em mais de 50% dos pares de cada nível relação CST.

Método	**Relação CST**					
	Identity	**Summary**	**Equivalence**	**Subsumption**	**Overlap**	**Permuta**
MLoc	5/5	3/4	4/6	5/8	5/8	9/14
MWol	5/5	1/4	2/6	4/8	4/8	4/14
MNol	5/5	2/4	3/6	5/8	4/8	1/14
MVol	5/5	2/4	3/6	5/8	2/8	2/14
MPdMorf	5/5	1/4	3/6	3/8	6/8	9/14
MSuj	5/5	1/4	3/6	2/8	2/8	0/14
MVp	5/5	1/4	5/6	1/8	0/8	0/14
MObPredp	5/5	1/4	0/6	2/8	1/8	0/14
MSin	5/5	0/4	4/6	1/8	3/8	1/14
MEtMorf	5/5	3/4	3/6	5/8	6/8	8/14

Tabela 4: Métodos *vs* níveis de redundância.

O maior número de relações CST (5 relações + "permuta"), em relação ao número de níveis de redundância (2 níveis + "nulo"), dificultou a interpretação humana (manual) dos dados da Tabela 4. Assim, as observações sobre a correlação entre os métodos e as relações CST são feitas a partir dos dados obtidos pelos algoritmos de AM, os quais são apresentados na sequência.

b) Análise automática

O algoritmo PART identificou o conjunto de 6 regras sequenciais descritas em (3), com 88,88% de precisão. A precisão mais baixa em relação à correlação entre métodos e redundância se deve pelo número maior de classes ou categorias a observar. Em outras palavras, os algoritmos de AM apresentaram menor precisão porque tiveram de lidar com um grupo maior de categorias nas quais os pares de sentenças deveriam ser classificados.

(3) **PART - Relações**

Se MNol <= 0.09 então **Permuta** (14.0)

Senão se MWol <= 0.8 e MVp = não e MEtMorf <= 0.9 e MLoc <= 0.36 e

MPdMorf <= 0.5 então **Subsumption** (10.0/3.0)

Senão se MVp = não AND MPdMorf > 0.33 então **Overlap** (7.0/1.0)

Senão se MWol <= 0.8 AND MVp = sim então **Equivalence** (6.0/1.0)

Senão se MPdMorf > 0.5 então **Identidade** (5.0)

Senão **Sumário** (3.0)

O algoritmo J48 gerou o conjunto de regras descritas em (4), com a mesma precisão que o PART, 88.88%. Em tais regras, observa-se o mesmo conjunto de métodos considerados pelo PART (MNol, MWol, MVp, MEtMorf, MLoc e MPdMorf).

(4) **J84 - Relações**

MNol <= 0.09 então **Permuta** (14.0)
MNol > 0.09
| MWol <= 0.8
| | MVp = sim então **Equivalence** (6.0/1.0)
| | MVp = não
| | | MEtMorf <= 0.9
| | | | MLoc <= 0.36
| | | | | MPdMorf <= 0.5 então **Subsumption** (10.0/3.0)
| | | | | MPdMorf > 0.5 então **Overlap** (6.0)
| | | | MLoc > 0.36 então **Sumário** (2.0/1.0)
| | | MEtMorf > 0.9 então **Sumário** (2.0)
| MWol > 0.8 então **Identidade** (5.0)

Esse algoritmo permite que as regras sejam visualizadas em árvore. No caso, a árvore relativa às regras do J48 para "métodos vs relações CST" neste trabalho é a da Figura 3.

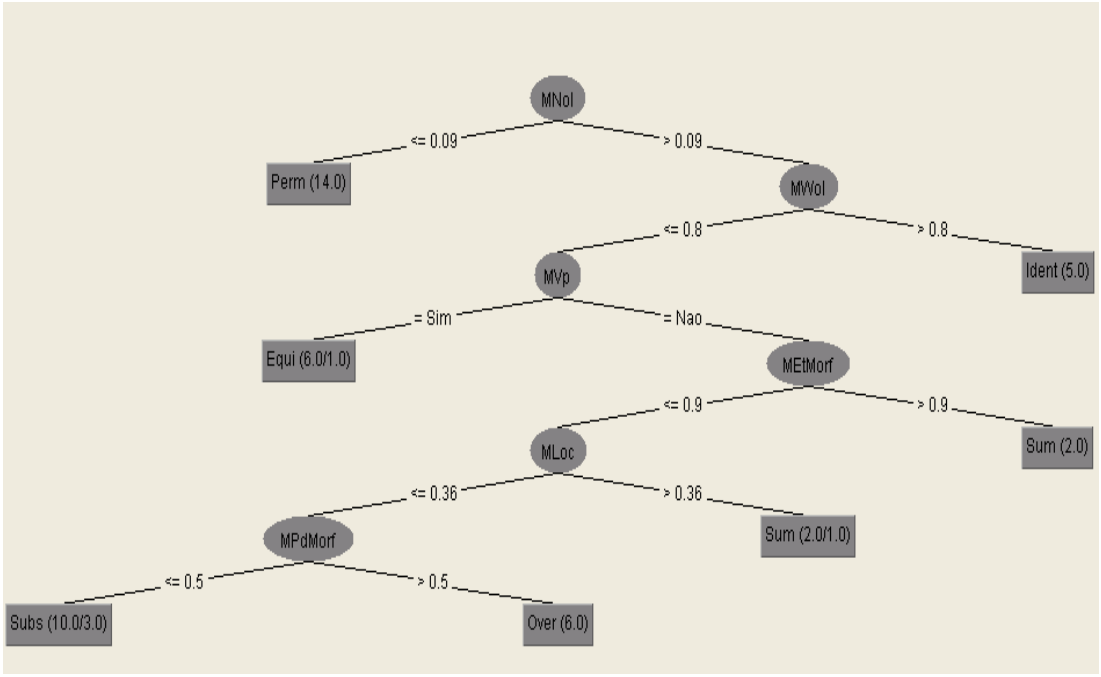


Figura 3: Visualização gráficas das regras geradas pelo J48.

4. Considerações Finais

Do estudo realizado neste trabalho, algumas observações são pertinentes. A primeira delas diz respeito ao fato de que o conjunto de dados de análise é pequeno, já que é composto por 45 pares de sentenças. Assim, os resultados obtidos sobre os métodos superficiais (estatísticos e linguísticos) de detecção da redundância na correlação com os níveis de redundância e as relações CST são indícios preliminares a respeito dos métodos, os quais poderão ser validados pela utilização das regras em uma aplicação de PLN, por exemplo, sumarização automática. A segunda é sobre a análise dos dados, que aqui seguiu duas estratégias, uma manual e outra automática, consideradas complementares. A terceira observação reside no fato de que este trabalho gerou um conjunto de regras de detecção da redundância entre pares de sentenças em português. Esse conjunto poderá servir de base para sistemas de PLN, especificamente os de sumarização multidocumento que, a partir de uma coleção de textos sobre um mesmo assunto, precisam selecionar a informações relevantes para compor um sumário sem redundância.

A quarta observação diz respeito às regras identificadas propriamente ditas. Na correlação entre os métodos e os níveis de redundância, observou-se que o método MNol é capaz de identificar pares que não são redundantes dos que apresentam alguma redundância e a redundância identificada como padrão foi a "parcial". O algoritmo PART precisa apenas de 4 atributos (métodos) (MNol, MVp, MEtMorf e MLoc) para identificar a redundância entre as sentenças com 97.7% precisão, ao passo que o J48 precisa de 5 atributos (MNol, MVp, MEtMorf, MLoc e MPdMorf). Na correlação entre os métodos e as relações CST, observa-se que MNol é capaz de identificar com precisão as sentenças que não estão relacionadas por relações CST. E, no caso, a relação identificada como padrão foi "Summary".

Por fim, enfatiza-se que a detecção da redundância continua sendo investigada pelo aluno em seu Trabalho de Conclusão de Curso (TCC). O objetivo do aluno no TCC é o de

estudar métodos de detecção da redundância pautados em outros atributos linguísticos (mais profundos) como a ocorrência de hiponímia ou de entidades nomeadas entre as sentenças de um par.

Referências Bibliográficas

ALEIXO, P.; PARDO, T.A.S. CSTNews: um corpus de textos jornalísticos anotados segundo a Teoria Discursiva Multidocumento CST (*Cross-document Structure Theory*). **Série de Relatórios Técnicos do ICMC**, São Carlos-SP, n. 326, 12p., 2008.

BRANCO, A; SILVA, J. Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 4, 2004, Lisbon. **Proceedings…** Lisbon, 2004, p. 507-510.

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. CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: RST BRAZILIAN MEETING, 3, 2001, Cuiabá, Brasil. **Proceedings…** Cuiabá, 2004, p. 88-105.

FELLBAUM, C. (Ed.). **WordNet: an electronic lexical database**. Ca, MA: MIT Press, 1998.

FERREIRA, A. B. H. **Novo dicionário eletrônico Aurélio da língua portuguesa**. Curitiba: Ed. Positivo, 2004. 1 CD-ROM

FREGE, G. Lógica e filosofia da linguagem. Tradução: Paulo Alcoforado. São Paulo: Cultrix/Edusp, 1978.

FRANK, E. WITTEN, I. H. HALL, M.A. **Data Mining: Pratical Machine Learning Toos and Techniques**. 3a Ed. MK. Waikato, 2011.

HATZIVASSILOGLOU, V.; KLAVANS, J. L.; ESKIN, E. Detecting text similarity over short passages: exploring linguistic feature combinations via Machine Learning. In: JOINT SIGDAT CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND VERY LARGE CORPORA, 1999, College Park, Maryland. **Proceedings…** Maryland, 1999, p. 203-12.

___ et al. SIMFINDER: A Flexible Clustering Tool for Summarization. In: NAACL WORKSHOP ON AUTOMATIC SUMMARIZATION, 2001, Pittsburg (PA), USA. **Proceedings…** Pittsburg, USA, 2001.

HENDRICKX, I.; DAELEMANS, W.; MARSI, E., KRAHMER, E. Reducing redundancy in multi-document summarization using lexical semantic similarity. In: WORKSHOP ON LANGUAGE GENERATION AND SUMMARISATION, 2009, Singapore. **Proceedings…** Singapore, 2009, p. 63–66.

HOUAISS, A.; VILLAR, M. de S. **Dicionário eletrônico Houaiss da língua portuguesa**. (versão 1.0). Rio de Janeiro: Editora Objetiva, 2001. 1 CD-ROM.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall: New Jersey, 2009.

LAGE, N. **Estrutura da notícia**. 5ª ed. São Paulo: Ática, 2002.

NEWMAN, E.; DOMN, W.; STOKES, N.; CARTHY, J.; DUNNION, J. Comparing redundancy removal techniques for multi-document summarization. In: STARTING AI RESEARCHERS' SYMPOSIUM, 2, 2004, Valencia. **Proceedings..**Valencia, 2004, p. 223-28.

MCKEOWN, K.; RADEV, D.R. Generating summaries of multiple news articles. In: INTERNATIONAL ACM-SIGIR, 18, 1995, Seattle. **Proceedings…**Seattle, 1995, p. 74-82.

MANI, I. **Automatic Summarization**. John Benjamins Publishing Co., Amsterdam, 2001.

___.; MAYBURY, M.T. **Advances in automatic text summarization**. The MIT Press, Cambridge, MA. 1999.

MARTINS, C. B. et al. Introdução à Sumarização Automática. **Rel. Técnico RT-DC 002/2001**, Departamento de Computação, UFSCar, São Carlos. Abril, 2001. 38p.

MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. Identifying Multidocument Relations. In: NLPCS, 7, 2010, Funchal, PT. **Proceedings...** Funchal, 2010, p. 60-69.

___.; PARDO, T.A.S.; DI FELIPPO, A.; DIAS-DA-SILVA, B.C. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TIL) 6, 2008, Vila Velha-ES, Brasil. **Proceedings...**2008, p. 390-392.

RADEV, D. R. et al. MEAD-a platform for multidocument multilingual text summarization. In: International Conference on Language Resources and Evaluation (LREC), 4, 2004, Lisbon. **Proceedings…** Lisbon, 2004, **Proceeings…** p. 1-4.

___. A common theory of information fusion from multiple text sources, step one: cross-document structure". In: ACL Signal Workshop on Discourse and Dialogue, 1, 2000, Hong Kong, **Proceedings…** Hong Kong, 2000, p. 74-83.

SINCLAIR, J. Corpus and text: basic principles. In: Wynne, M. (Ed.). **Developing linguistic corpora: a guide to good practice**. Oxford: Oxbow Books, 2005. p.1-16. Disponível em: <www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm>. Acesso em: 02 ago. 2010.

SPARCK JONES, K. Discourse modeling for Automatic Summarization. **Tech. Report No. 290**. University of Cambridge. UK, February, 1993.

____. Automatic summarising: a review and discussion of the state of the art. **Technical Report UCAM-CL-TR-679**. University of Cambridge. 2007.]

APÊNDICE 1

Par	Relação	*Cluster*	Sentenças originais
1	Identity	D2_C1	As vítimas do acidente foram 14 passageiros e três membros da tripulação.
		D3_C1	As vítimas do acidente foram 14 passageiros e três membros da tripulação.
2	Identity	D2_C1	Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.
		D3_C1	Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.
3	Identity	D3_C10	Comandos israelenses mataram outros três guerrilheiros libaneses na cidade de Tiro, onde destruíram sete plataformas de lançamento de foguetes, informaram as fontes israelenses.
		D4_C10	Comandos israelenses mataram outros três guerrilheiros libaneses na cidade de Tiro, onde destruíram sete plataformas de lançamento de foguetes, informaram as fontes israelenses.
4	Identity	D3_C11	Na capital, houve ataques a outros quatro ônibus.
		D4_C11	Na capital, houve ataques a outros quatro ônibus.
5	Identity	D1_C12	Os crimes aconteceram na cidade de Muttur, que durante as últimas duas semanas vive graves conflitos entre as tropas do Exército do Sri Lanka e a guerrilha dos Tigres de Libertação da Pátria Tâmil (LTTE).
		D2_C12	Os crimes aconteceram na cidade de Muttur, que durante as últimas duas semanas vive graves conflitos entre as tropas do Exército do Sri Lanka e a guerrilha dos Tigres de Libertação da Pátria Tâmil (LTTE).
6	Sumary	D2_C6	Lula disse que o critério para o investimento nas cidades será técnico, não partidário.
		D1_C6	"O critério é eminentemente técnico, ou seja, eu não quero saber se o prefeito é do PFL, do PT, do PMDB, do PSDB, do PTB, do PR, do PC do B.
7	Sumary	D2_C21	O cronograma da obra depende de estudos finais da Infraero.
		D3_C21	O cronograma da obra depende de estudos finais que estão sendo realizados pela Infraero.
8	Sumary	D4_C11	Até o momento, não há registro de feridos.
		D5_C11	Os ataques desta madrugada, até agora, não deixaram mortos ou feridos.
9	Sumary	D3_C21	O ministro da Defesa, Nelson Jobim, decidiu que será realizada uma reforma definitiva na pista principal de Guarulhos, o mais rápido possível, de acordo com a assessoria do ministério da Defesa.
		D3_C21	Para receber os vôos de Cumbica, Viracopos precisará ser ampliado, sobretudo seu terminal de passageiros, segundo nota do Ministério da Defesa.
10	Equivalence	D1_C2	A margem de erro é de dois pontos percentuais, para mais ou para menos.
		D2_C2	A margem de erro é de 2 pontos porcentuais.
11	Equivalence	D1_C3	A TAM afirma que "o procedimento não configura qualquer obstáculo ao pouso da aeronave".
		D3_C3	De acordo com a empresa aérea, o reversor estava travado, mas argumentou que a aeronave tinha condições de pouso normais, mesmo sem ele.
12	Equivalence	D1_C7	"É um par de irmãos admirável, cada um com cerca de 1% da massa do Sol", disse Jayawardhana.
		D2_C7	"Este é um par de gêmeos verdadeiramente de destaque, já que cada um tem uma massa de apenas 1% de nosso Sol", declarou Jayawardhana.
13	Equivalence	D1_C9	A polícia também vai abrir nova investigação sobre a participação de desembargadores e conselheiros do Tribunal de Contas no suposto esquema.
		D2_C9	A PF abriu uma nova frente de atuação para apurar o caso com o objetivo de apurar a participação de desembargadores e conselheiros do Tribunal de Contas na quadrilha.
14	Equivalence	D2_C15	Nove pessoas morreram, três delas crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão ocorrida em um mercado de Moscou, informou a polícia.
		D4_C15	MOSCOU (Rússia) - Nove pessoas morreram, sendo três crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão registrada em um mercado moscovita, informou a Polícia de Moscou.
15	Equivalence	D1_C16	O prazo foi definido pela Mesa Diretora da Câmara.
		D3_C16	O prazo foi definido pela direção da Câmara.
16	Subsumption	D3_C8	Com o resultado, o Brasil continua sendo a única equipe invicta da competição, mantendo a liderança do Grupo B da Liga, com sete vitórias em sete partidas.
		D1_C8	Invicto na competição, o Brasil está tranquilo na liderança do Grupo B.
17	Subsumption	D3_C10	A aviação de Israel realizou durante a madrugada desta segunda-feira, dia 7, ataques a 150 alvos no Líbano.
		D5_C10	A Força Aérea israelense lançou uma série de bombardeios contra o Líbano nesta segunda-feira.
18	Subsumption	D5_C11	Calcula-se em 15 o número de ônibus incendiados, sendo dez na região do ABC e quatro na capital

		D4_C11	Na capital, houve ataques a outros quatro ônibus.
19	Subsumption	D2_C12	"Tentamos enviar uma equipe a Muttur para averiguar o que está acontecendo, mas os soldados não permitiram que entrássemos na cidade, que está totalmente bloqueada", afirmou.
		D3_C12	Segundo ele, o grupo pretendia enviar uma equipe à região, mas foi impedido por soldados.
20	Subsumption	D1_C13	Quinze voluntários da ONG francesa Ação Contra a Fome (ACF) foram assassinados no nordeste do Sri Lanka, informou hoje um porta-voz da organização.
		D3_C13	Quinze funcionários locais de uma organização de caridade francesa no Sri Lanka foram encontrados mortos na cidade de Muttur, no norte do país.
21	Subsumption	D4_C14	CAIRO - Pelo menos 80 pessoas morreram e mais de 165 ficaram feridas nesta segunda-feira após a colisão de dois trens de passageiros no delta do Nilo, ao norte do Cairo, informaram fontes policiais e médicas.
		D2_C14	Fontes policiais e médicas informaram anteriormente que pelo menos 80 pessoas tinham morrido no acidente.
22	Subsumption	D2_C15	Nove pessoas morreram, três delas crianças, e outras 25 ficaram feridas nesta segunda-feira em uma explosão ocorrida em um mercado de Moscou, informou a polícia.
		D3_C15	Anteriormente, a Polícia havia informado sobre nove mortos, sendo três deles crianças, e 25 feridos.
23	Subsumption	D2_C16	O horário-limite para que o parlamentar renuncie - 20 horas - foi estabelecido pela direção da Câmara a fim de que o ato seja oficializado com a sua publicação já no Diário Oficial do Congresso de amanhã.
		D1_C16	O prazo foi definido pela Mesa Diretora da Câmara.
24	Overlap	D1_C17	Alckmin acusou Lula de arrogante, de subestimar a inteligência dos brasileiros e relacionou o presidente aos escândalos do mensalão, sanguessuga e ao caso Waldomiro Diniz.
		D2_C17	"Ele [Lula] trabalhou ao lado do Waldomiro [Diniz], do mensalão, dos sanguessugas, de todos esses escândalos. Isto que é o fato, isto que é grave".
25	Overlap	D1_C18	Um atirador matou ao menos 30 pessoas em dois diferentes locais da Universidade Técnica da Virgínia, em Blacksburg (Virgínia), nesta segunda-feira, no pior ataque a tiros contra um campus universitário da história dos Estados Unidos.
		D3_C18	A Casa Branca em nota divulgada pela imprensa local considerou o atentado, o pior ataque a tiros um campus universitário da história dos Estados Unidos.
26	Overlap	D1_C19	O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque.
		D2_C19	Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido à hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe.
27	Overlap	D1_C20	No entanto, deputados de oposição tentam obstruir a votação para adiar a aprovação da prorrogação da PEC da CPMF.
		D4_C20	O requerimento faz parte da estratégia dos partidos de oposição de adiar a votação da matéria.
28	Overlap	D1_C21	SÃO PAULO - A pista principal do Aeroporto Internacional de São Paulo (Cumbica), em Guarulhos, será totalmente reformada em março de 2008, segundo informações do Ministério da Defesa anunciadas nesta segunda-feira, 6.
		D2_C21	O Ministério da Defesa anunciou nesta segunda-feira (6) que em março do ano que vem uma das pistas do Aeroporto de Guarulhos será fechada para reformas de seu trecho central.
29	Overlap	D2_C22	A TAM cancelou 68 vôos previstos para esta terça em Congonhas e transferiu outros 22 para Cumbica.
		D5_C22	A TAM anunciou o cancelamento de 68 vôos nesta terça (24) e o remanejamento de outros 22.
30	Overlap	D1_C23	As piores enchentes dos últimos 60 anos no Reino Unido deixaram milhares de britânicos desabrigados, sem abastecimento de água ou sem energia elétrica.
		D2_C23	LONDRES - A chuva torrencial que atinge o Reino Unido encobriu estradas e milhares de pessoas estão sem fornecimento de eletricidade e de água potável em decorrência da pior enchente nos últimos 60 anos no país, segundo informou nesta segunda-feira, 23, a rede de televisão BBC.
31	Overlap	D3_C24	RIO - Depois da queda de April Steiner, a brasileira Fabiana Murer leva a medalha de ouro no salto com vara, com 4m50 - novo recorde pan-americano.
		D4_C24	A brasileira Fabiana Murer conquistou o primeiro ouro do atletismo para o Brasil, nesta segunda-feira, na prova de salto com vara.
32	Permuta	D4_C25	As seleções de vôlei e futebol conquistaram a Liga Mundial e a Copa América e escreveram mais uma vez o nome do Brasil nos respectivos esportes.
		D4_C26	KINGSTON - O furacão Dean chegou à costa sul da Jamaica, inundando a capital e espalhando árvores e telhados depois de matar nove pessoas na passagem pelo Caribe nesta segunda-feira, na direção da península de Yucatán, no México.
33	Permuta	D2_C27	O Brasil lavou a alma após o decepcionante empate com a Colômbia no último domingo e, nesta quarta-feira, aplicou uma sonora goleada por 5 a 0 sobre o Equador no Maracanã.
		D3_C28	KATOWICE (Polônia) - A seleção brasileira de vôlei voltou a fazer bonito, desta vez na final da Liga Mundial, disputada contra a Rússia neste domingo no ginásio de Spodekna, em Katowice, na Polônia.
34	Permuta	D1_C29	Desde 2002, mais de mil pessoas deram entrada em processos contra a Igreja Católica por abusos sexuais na Califórnia e, nos últimos anos, a arquidiocese de Los Angeles já pagou US$ 114 milhões a 86 vítimas.
		D3_C30	SÃO PAULO - O Itaú obteve nos primeiros seis meses deste ano o maior lucro já registrado por um banco privado do país nos últimos vinte anos.

35	Permuta	D1_31	A tocha passará por vinte países, mas o Brasil não estará no percurso olímpico.
		D2_C32	TÓQUIO - Os chefes da usina nuclear do Japão atingida por terremotos na última segunda-feira admitiram que ocorreram mais vazamentos radioativos no local.
36	Permuta	D4_C33	O Brasil sediou a conferencia das Nações Unidas para o Desenvolvimento e Meio-ambiente, a Rio 92. Precisamos avaliar o caminho percorrido e estabelecer novas metas. Proponho em 2012 uma nova conferência que o Rio se propõe a sediar, a Rio +20.
		D2_C24	Hoje, se a pessoa cai na malha fina, a Receita fiscaliza também as declarações de anos anteriores.
37	Permuta	D5_35	Um dos traficantes internacionais de drogas mais procurados pelos Estados Unidos, o colombiano Juan Carlos Ramirez-Abadia, de 44 anos, foi preso por volta das 6h30 desta terça-feira (7), na região de Aldeia da Serra, na Grande São Paulo.
		D3_C36	BRASÍLIA e SÃO PAULO - Foi confirmada às 11h40m desta sexta-feira a morte do senador Antônio Carlos Magalhães (DEM-BA), de 79 anos.
38	Permuta	D1_37	A cadeia abriga 203 detentos, mas só tem capacidade para 80.
		D5_C41	O recorde anterior era de Fernando Scherer com quatro em Winnipeg, em 99.
39	Permuta	D2_C45	Junto com Marinho, policiais civis detiveram um grafiteiro de 27 anos que tem passagem na polícia por roubo.
		D3_C46	A agência meteorológica do Japão chegou a emitir um alerta de tsunami para a Ilha Sado, na costa da província de Niigata, mas suspendeu o aviso uma hora depois.
40	Permuta	D1_C48	Quando o nome do treinador foi anunciado, parte dos torcedores também o vaiou.
		D4_C50	Para salários acima desse limite, haverá um desconto fixo no valor de 214 reais, também na declaração do imposto de renda.
41	Permuta	D3_C39	Os mandados de busca e apreensão foram expedidos pela juíza federal Maria Isabel do Prado, da 2ª Vara Federal de Guarulhos.
		D4_C39	Depois das vaias, Lula desistiu de declarar aberto os jogos, como estava planejado.
42	Permuta	D3_C49	A declaração foi feita pelo presidente do Comitê Olímpico Brasileiro (COB), Carlos Nuzman.
		D3_C36	Foi confirmada às 11h40m desta sexta-feira a morte do senador Antônio Carlos Magalhães (DEM-BA), de 79 anos.
43	Permuta	D1_C18	Um deles foi morto em um dormitório e outros foram assassinados dentro de uma sala de aula, segundo o chefe de polícia do campus, W. R. Flinchum.
		D3_C21	Mas essa ampliação deverá estar concluída até março, sendo feita simultaneamente à reforma de Guarulhos.
44	Permuta	D1_C36	No final de maio, o parlamentar sentiu-se mal no Senado e chegou a cair em frente ao seu gabinete.
		D5_C41	O recorde anterior era de Fernando Scherer com quatro em Winnipeg, em 99.
45	Permuta	D1_C48	Quando o nome do treinador foi anunciado, parte dos torcedores também o vaiou.
		D4_C50	Para salários acima desse limite, haverá um desconto fixo no valor de 214 reais, também na declaração do imposto de renda.