Programa de Pós-Graduação em Lingüística

PPGL

INVESTIGAÇÃO DO FENÔMENO DA COMPLEMENTARIDADE PARA A SUMARIZAÇÃO AUTOMÁTICA MULTIDOCUMENTO

JACKSON WILKE DA CRUZ SOUZA

SÃO CARLOS

2014

Universidade Federal de São Carlos

RESUMO

A Sumarização Automática Multidocumento (SAM) é uma alternativa computacional para o tratamento da grande quantidade de informação disponível on-line. Nela, busca-se gerar automaticamente um único sumário coerente e coeso a partir de uma coleção de textos que tratam de um mesmo assunto, sendo cada um deles proveniente de uma fonte distinta. Para tanto, a SAM precisa selecionar a informação mais importante da coleção para compor o sumário. A seleção do conteúdo principal requer a identificação da redundância, complementaridade e contradição, que se caracterizam por serem os fenômenos multidocumento existentes entre os textos-fonte do sumário. Acerca da complementaridade, em especial, a identificação é relevante, pois uma informação pode ser selecionada para o sumário uma vez que completa outra já selecionada. Em vários métodos de SAM, os fenômenos multidocumento são identificados por meio da análise automática dos textos-fonte com base na CST (Cross-document Structure Theory). Entretanto, para o Português do Brasil (PB) os métodos que são empregados na identificação, sobretudo, automática do fenômeno da complementaridade são baseados em pouco conhecimento linguístico, restringindo-se, por vezes, a conhecimento estatístico. O saber linguístico empregado restringe-se à similaridade lexical que exista entre o par de sentenças; dada a similaridade, calcula-se a possibilidade de um par de sentenças possuírem uma das relações de complementaridade, por exemplo. Dessa forma, propõe-se investigar o fenômeno da complementaridade em um corpus multidocumento com o objetivo de identificar as características linguísticas de seus diferentes tipos, as quais poderão subsidiar a identificação automática das relações CST de complementaridade. Assim, enrique-se ia o cenário de identificação das relações do modelo CST, o qual, para o PB, conta, sobretudo, com conhecimento de similaridade (ou redundância). As características do fenômeno da complementaridade partem da descrição linguística do fenômeno. Com base em corpus, descreve-se quais atributos codificam o fenômeno, com a finalidade de automatizar esse processo. Além disso, a descrição linguística com base em corpus permitirá compreender a razão pela qual por vezes o fenômeno pode não ser modelado automaticamente. A fim de enriquecer a pesquisa e a análise dos dados, utilizar-se-á algoritmos de aprendizado de máquina capazes de ampliar o olhar sobre o fenômeno, com base na análise e descrição manual .Os resultados dessa análise podem ser comparados quanto ao desempenho manual e automático, além de estabelecer um comparativo com o cenário atual de identificação das relações CST em PB.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1. INTRODUÇÃO

1.1. Contextualização

Na subárea do Processamento Automático das Línguas Naturais (PLN) denominada Sumarização Automática Multidocumento (SAM), objetiva-se automatizar a produção de sumários a partir de uma coleção de textos-fonte, advindos de fontes distintas, que abordam um mesmo assunto (MCKEOWN; RADEV, 1995). As pesquisas sobre SAM têm sido motivadas principalmente pela enorme quantidade de informação textual disponível na web e o pouco tempo que se tem para assimilá-la (MANI, 2001).

Tais pesquisas têm visado majoritariamente à produção de sumários extrativos (ou extratos) (ou seja, sumários compostos comumente por sentenças copiadas integralmente dos textos-fonte) que sejam informativos (isto é, veiculam o conteúdo central da coleção, substituindo a leitura dos textos-fonte) e genéricos (ou seja, voltados para uma audiência não específica) (KUMAN, SALIM, 2012).

Pensando-se nos processos básicos que compõem a Sumarização Automática (SA), os referidos sumários multidocumento têm sido gerados em 3 etapas: (i) análise, (ii) transformação e (iii) síntese (SPARCK-JONES, 1993; MANI, 2001) (

Figura 1).

Figura 1 – Arquitetura genérica de um sistema de SAM.

Fonte – Sparck Jones, 1993

Na análise, os textos-fonte são interpretados, extraindo-se uma representação formal dos mesmos. A transformação é a etapa principal, pois, a partir da representação gerada na análise, o conteúdo dos textos-fonte é condensado em uma representação interna do sumário. Essa condensação é resultante da seleção de conteúdo, que consiste em ranquear os segmentos dos textos-fonte (comumente, sentenças) em função de algum

critério de relevância e selecionar os de maior pontuação para compor o sumário até que a taxa de compressão (o tamanho desejado do sumário) seja atingida. Na síntese, produz-se o sumário em língua natural a partir do conteúdo selecionado.

A complexidade dessas três etapas depende diretamente da abordagem ou paradigma de sumarização empregado. De acordo com a quantidade e o nível de conhecimento linguístico, a SAM pode ser superficial ou profunda (MANI, 2001).

Os métodos/sistemas superficiais realizam a SAM com base em pouco ou nenhum conhecimento linguístico, pois o tratamento dos textos-fonte pauta-se comumente em dados estatísticos. Por essa razão, esses métodos/sistemas geram extratos, com baixo custo de desenvolvimento, robustez e escalabilidade. Por outro lado, eles produzem sumários menos coerentes, coesos e informativos.

Os métodos/sistemas profundos usam conhecimento linguístico codificado em gramáticas, repositórios semânticos e modelos de discurso. Assim, o desenvolvimento dos métodos/sistemas é caro e sua aplicação é mais restrita. O desempenho, no entanto, é superior, pois os sumários são mais coerentes, coesos e informativos, podendo ser extrativos ou abstrativos (isto é, produzidos pela reescrita dos textos-fonte).

Vários métodos/sistemas superficiais e profundos têm sido desenvolvidos para produzir extratos informativos e genéricos a partir de coleções de textos do gênero jornalístico (KUMAR, SALIM, 2012).

Tendo em vista a produção desse tipo de sumário (extrativo, informativo e genérico), é preciso selecionar as sentenças mais importantes de uma coleção de textos para compor o seu respectivo sumário, evitando-se, por um lado, que ele seja formado por sentenças redundantes e contraditórias e permitindo, por outro lado, que também veicule sentenças com conteúdo complementar. Em outras palavras, a necessidade de identificação dos fenômenos multidocumento ocorre por quê: (i) as sentenças mais redundantes na coleção veiculam suas principais informações e, por isso, devem constar no sumário; (ii) as sentenças com conteúdo complementar já selecionadas podem compor o sumário, e (iii) as sentenças redundantes ou contraditórias entre si não devem ser selecionadas para o sumário.

Por conseguinte, é preciso identificar, na fase de análise[1], os fenômenos de conteúdo típicos da multiplicidade de textos-fonte, sobretudo os jornalísticos. Tais

[1] Em um método/sistema de SAM, a análise é feita por um analisador automático discursivo, como o CSTParser (MAZIERO, PARDO, 2011), desenvolvido para o português do Brasil (PB).

fenômenos são a redundância, a complementaridade e a contradição, os quais são ilustrados pelos pares de sentenças em (1), (2) e (3), respectivamente.

(1) Sentença 1: A margem de erro é de dois pontos percentuais, para mais ou para menos.
Sentença 2: A margem de erro é de 2 pontos porcentuais.

(2) Sentença 1: Em Niigata, um terremoto em outubro de 2004, também de magnitude 6,8, matou 65 pessoas e deixou mais de 3.000 feridos.
Sentença 2: No caso do Japão, a magnitude apontada de 6,8 é considerada "forte".

(3) Sentença 1: José Maria Eymael, do PSDC, e Rui Pimenta, do PCO, não chegaram a obter 1% das intenções de voto.
Sentença 2: Os candidatos José Maria Eymael (PSDC) e Ruy Pimenta (PCO) não pontuaram.

Entre as sentenças de (1), por exemplo, há uma relação de redundância, já que ambas expressam o mesmo conteúdo, ainda que por meio de paráfrase. Entre as sentenças de (2), por sua vez, há uma relação de complementaridade, já que a Sentença 2 detalha uma informação contida na Sentença 1 ("*a magnitude apontada de 6,8 é considerada 'forte'*"). E, finalmente, entre as sentenças de (3), observa-se uma relação de contradição, uma vez que ambas as sentenças não expressam a mesma informação (a Sentença 2 aponta que os candidatos em questão não pontuaram, e na Sentença 1 os mesmos candidatos obtiveram pontuações muito baixas).

Na literatura, encontram-se várias propostas para a análise multidocumento, ou seja, para o relacionamento de segmentos textuais advindos de documentos distintos sobre mesmo assunto (p.ex.: TRIGG, 1983, TRIGG; WEISER, 1986, ALLAN, 1996, RADEV; MACKEOWN, 1998, AFANTENOS et al., 2004 e DAGAN et al., 2005). A Figura 2 ilustra um esquema genérico de relacionamento entre sentenças de textos que abordam mesmo tópico.

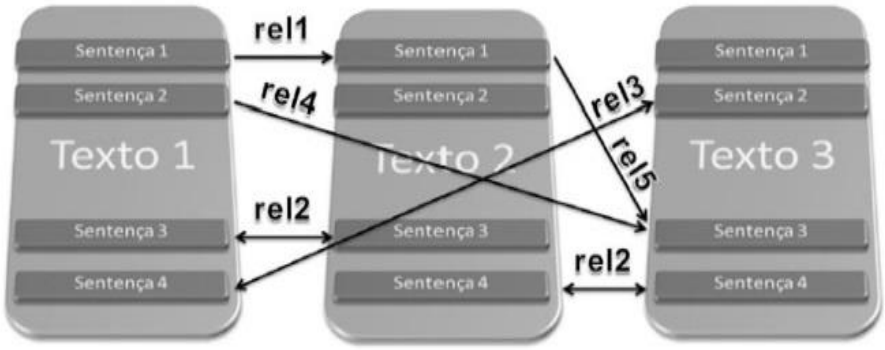

Figura 2 - Esquema genérico de análise multidocumento.

Fonte: Maziero (2012).

No cenário da SAM, a análise multidocumento em vários métodos/sistemas profundos baseia-se em conectar (em pares) sentenças de textos distintos de uma coleção pelas relações da teoria/modelo *Cross-document Structure Theory* (CST) (RADEV, 2000).

No Quadro 1, tem-se o conjunto refinado para o Português do Brasil (PB) de relações CST de Maziero et al (2010), elaborado a partir da anotação manual do *corpus* denominado CSTNews (CARDOSO et al., 2011).

Quadro 1- Conjunto refinado de relações CST de Maziero et al. (2010).

Identity	*Elaboration*
Equivalence	*Contradiction*
Summary	*Citation*
Subsumption	*Attribution*
Overlap	*Modality*
Historical background	*Indirect speech*
Follow-up	*Translation*

Fonte: Maziero et al. (2010).

De acordo com uma tipologia proposta por Maziero et al. (2010), algumas relações CST capturam a "complementaridade" entre sentenças de um par. De um modo geral, entende-se complementaridade como a relação que se estabelece entre duas sentenças, S1 e S2, sendo cada uma delas proveniente de um texto distinto, quando S2 apresenta informação complementar (ou seja, adicional ou suplementar) ao conteúdo veiculado

por S1. Assim, S1 e S2 possuem conteúdo em comum, sendo que S2 apresenta informação aditiva não prevista em S1.

Para evidenciar a relevância desse fenômeno em um *corpus* multidocumento, ressalta-se que, no CSTNews, há 713 pares de sentenças com relações de complementaridade, de um total de 1650 pares anotados manualmente com base na CST, o que equivale a 43% das relações.

Ainda de acordo com a tipologia de Maziero et al. (2010), as relações de complementaridade podem ser temporais ou atemporais. As temporais podem ser de dois tipos diferentes. Dado um par de sentenças S1 e S2, as mesmas são complementares do subtipo temporal quando: (i) S2 apresenta informações históricas/ passadas sobre algum elemento presente em S1 (no modelo CST, essa relação é rotulada como *historical background*); (ii) S2 apresenta acontecimentos/ eventos que sucederam os acontecimentos/ eventos presentes em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si (no modelo CST, essa relação é rotulada como *follow-up*).

A relação CST atemporal não envolve conteúdo que indica a localização no tempo (anterior ou posterior) de um acontecimento/fato em relação a outro. Ela se estabelece quando, dado um par de sentenças S1 e S2, S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1. Além disso, o elemento elaborado em S2 deve ser o foco de S2. No modelo CST, essa relação é rotulada por *elaboration.*

A partir da identificação das relações CST na SAM, as sentenças são pontuadas e ranqueadas em função do número de relações que possuem na coleção (p.ex.: RADEV, MCKEOWN, 1998; ZHANG et al., 2002). Assim, considerando-se o número de relações CST como critério de relevância, as sentenças mais conectadas, que ocupam o topo do ranque, são selecionadas para o sumário porque veiculam as informações principais da coleção. Além disso, o tipo das relações também pode ser utilizado para selecionar conteúdo a compor o sumário. Por exemplo, caso uma sentença do ranque, candidata a compor o sumário, esteja em relação de complementaridade com outra já selecionada, esta pode vir a compor o sumário caso não ultrapasse a taxa de compressão.

Segundo Zhang e Radev (2005), as relações CST se dão entre sentenças que possuem algum tipo de sobreposição de conteúdo e/ou forma. Por essa razão, a identificação automática das relações CST de conteúdo das sentenças (inclusive a

complementaridade) tem sido feita com relativo sucesso, baseando-se quase que exclusivamente na similaridade lexical existente entre 2 sentenças.

A similaridade é modelada por um conjunto de atributos (p.ex.: sobreposição de palavras de conteúdo) e capturada por medidas estatísticas (p.ex.: *word overlap*) que, mediante o valor obtido, indicam o fenômeno (redundância, complementaridade ou contradição) e a relação CST correspondente (p.ex.: ZHANG et al., 2003, ZHANG, RADEV, 2005, MAZIERO et al, 2010).

Para o PB, o CSTParser identifica as relações CST com precisão aproximada de 70%, baseando-se em atributos similares aos de Zhang et al. (2003) e Zhang e Radev (2005) e em algumas regras. Dentre os atributos, por exemplo, estão: (i) sobreposição de sequências de palavras; (ii) sobreposição de nomes próprios; (iii) sobreposição de numerais; (iv) ocorrência de palavras sinônimas, etc. (MAZIERO, 2012).

Para a identificação da similaridade, em especial, há outros atributos que podem ser utilizados, como: (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 etiquetas morfossintáticas, (vi) ocorrência de itens lexicais que compartilham mesmo hiperônimo, (vii) sobreposição de entidades mencionadas, etc. (HATZIVASSILOGLOU et al., 1999, 2000, NEWMAN et al., 2004, JIJKOUN; RIJKE, 2005, HENDRICKX et al., 2009, KUMAR et al., 2012, SOUZA et al., 2012, 2013 e SOUZA, 2013).

Do que foi exposto, observa-se que: (i) as relações CST de conteúdo, inclusive as de complementaridade, são identificadas em função de alguns atributos linguísticos que capturam apenas a similaridade entre duas sentenças, posto que sentenças complementares apresentam certo conteúdo redundante; (ii) há outros atributos na literatura por meio dos quais a redundância ou similaridade pode ser identificada, e (iii) não há atributos que traduzem características específicas da complementaridade, a não ser certa redundância.

Dessa forma, buscando melhorar a identificação automática da complementaridade, propõem-se os objetivos descritos na próxima subseção.

1.2. Objetivos e hipóteses

Neste trabalho, visa-se a investigação do fenômeno da complementaridade na SAM. De um modo geral, objetiva-se realizar uma descrição linguística desse fenômeno multidocumento e propor métodos de identificação automática do mesmo em PB.

Assim, os objetivos específicos são:

a) descrever as características linguísticas da complementaridade com base em *corpus*;
b) "traduzir" as características da complementaridade em atributos linguísticos (superficial e/ou profundo) capazes de distinguir os diferentes tipos de complementaridade (temporal e atemporal) e as relações CST que os codificam (*historical background*, *follow-up* e *elaboration*).

Tais objetivos, aliás, relacionam-se diretamente à meta de pesquisa de um projeto maior denominado SUSTENTO[2] (FAPESP 2012/13246-5/ CNPq 483231/2012-6)[3], que é a de produzir e/ou sistematizar conhecimento linguístico para subsidiar a SAM do PB.

Os objetivos deste trabalho pautam-se em 4 hipóteses iniciais sobre o fenômeno da complementaridade e da sua identificação automática no cenário multidocumento:

- **Hipótese 1:** atributos superficiais e profundos de detecção da redundância são pertinentes para a identificação da complementaridade, já que o conteúdo entre duas sentenças pode estar sob certa sobreposição em relação complementar.
- **Hipótese 2:** a complementaridade pode se manifestar na superfície linguística, e essa manifestação pode ser capturada por atributos específicos que tem o potencial de subsidiar métodos automáticos de detecção desse fenômeno.
- **Hipótese 3:** métodos de detecção da complementaridade podem capturar os diferentes tipos de complemento (temporais e atemporais).
- **Hipótese 4:** métodos de detecção da complementaridade capturam as relações CST que expressam complemento (*historical background*, *follow-up* e *elaboration*).

Visando alcançar os objetivos, estabeleceram-se as etapas metodológicas a seguir.

[2] Disponível em: http://www.nilc.icmc.usp.br/arianidf/sustento/
[3] O projeto SUSTENTO subsidia o projeto Sucinto (FAPESP 2012/03071-3) com pesquisas e trabalhos linguísticos. O Sucinto visa produzir recursos, ferramentas e sistemas de Sumarização Automática. Disponível em: http://www.icmc.usp.br/pessoas/taspardo/sucinto/

1.3. Metodologia

Equacionou-se metodologicamente esta pesquisa em 8 etapas ou tarefas. Salienta-se que a Tarefa 1 é contínua. As Tarefas 2, 3 e 4 já foram realizadas. As Tarefas 5, 6 e 7 encontram-se em andamento. A Tarefa 8 se trata de etapa futura.

Tarefa 1 – Revisão da literatura: consiste no estudo sobre a CST e delimitação dos fenômenos multidocumento de acordo com o modelo teórico, sobretudo o fenômeno da complementaridade. Além disso, investigação constante de métodos de identificação automática das relações CST ou modelos similares que estudam o relacionamento entre porções textuais.

Tarefa 2 – Seleção e recorte do *corpus*: consiste na seleção de um *corpus* multidocumento em PB cujos textos-fonte do gênero jornalístico tenham sido anotados via CST e na criação de *subcorpus* que proporcione o estudo da complementaridade. O *subcorpus* será composto basicamente por pares de sentenças relacionadas pelas relações CST de complementaridade e por pares de sentenças que apresentam complementaridade nula. Uma parcela do *subcorpus* será destinada à aplicação dos métodos para identificação dos mais eficientes (*corpus* de treinamento) e outra parcela será destinada ao teste dos métodos de detecção mais eficientes (*corpus* de teste).

Tarefa 3 – Análise de *corpus*: consiste na análise manual do *subcorpus* construído na Tarefa 2. Especificamente, os pares cujas sentenças estão conectadas pelas relações *historical background*, *follow-up* e *elabaration* serão analisados com o objetivo de identificar as características linguísticas da complementaridade e "traduzi-las" em atributos (superficiais e profundos) capazes de distinguir os diferentes tipos de complementaridade (temporal e atemporal) e as relações CST que os codificam.

Tarefa 4 – Caracterização de *corpus*: consiste na caracterização ou descrição manual e/ou semiautomática de cada uma das sentenças do *subcorpus* (com ou sem complementaridade) em função dos atributos identificados na Tarefa 3.

Tarefa 5 – Identificação de métodos de detecção da complementaridade: consiste em identificar, com base nas sentenças caracterizadas, os atributos mais relevantes para a identificação da complementaridade. A identificação dos atributos poderá ser manual e/ou automática. Quando automática, as sentenças caracterizadas são submetidas a um ambiente de Aprendizado de Máquina que aprende padrões estatisticamente relevantes, gerando regras que subsidiam métodos automáticos de detecção da complementaridade.

Tarefa 6 – Estudo da correlação entre os métodos e os tipos de relação CST: essa tarefa consiste no estudo da correlação entre os métodos delimitados na Tarefa 5 e os tipos de complementaridade (temporais e atemporais). Com isso, pretende-se identificar os métodos que expressam mais adequadamente as diferenças de complemento.

Tarefa 7 – Estudo da correlação entre os métodos e as relações CST: consiste no estudo da correlação entre os métodos delimitados na Tarefa 5 e as relações CST de *historical background*, *follow-up e elabaration*. Com isso, pretende-se identificar os métodos que expressam mais adequadamente as relações.

Tarefa 8 – Avaliação: consiste na aplicação dos métodos mais eficientes identificados nas Tarefas 6 e 7 à uma parcela do *subcorpus* distinta da utilizada na fase de treinamento (Tarefas 6 e 7), ou seja, à parcela de teste. Além disso, comparar a identificação da complementaridade por métodos manuais e automáticos (por aprendizado de máquina, por exemplo).

1.4. Cronograma

No Quadro 2, apresenta-se o cronograma para a realização desta pesquisa, o qual está dividido em trimestres e compreende o período de 2 anos, tendo iniciado em 03/2013.

Quadro 2- Cronograma original

	2013			2014				2015
Tarefas	**1 T.**	**2 T.**	**3 T.**	**4 T.**	**5 T.**	**6 T.**	**7 T.**	**8 T.**
Integralização dos créditos								
Tarefa 1								
Tarefa 2								
Tarefa 3								
Tarefa 4								
Tarefa 5								
Tarefa 6								
Redação e defesa da Qualificação								
Tarefa 7								
Tarefa 8								

Fonte: Elaborado pelo autor.

1.5. Estrutura da qualificação

Este texto está organizado em 5 seções. Na Seção 2, apresenta-se a revisão da literatura realizada até a escrita da qualificação. Como resultado dessa revisão, apresenta-se a CST e as relações de complementaridade, e descrevem-se os principais métodos para a identificação automática das relações CST. Na Seção 3, descrevem-se as tarefas realizadas, destacando a análise da complementaridade em *corpus* e a conseguinte delimitação de atributos linguísticos que a caracterizam. Além disso, elencam-se as tarefas em andamento e as futuras. Na Seção 4, apresenta-se o cronograma atualizado. Na Seção 5, tecem-se comentários finais sobre o trabalho.

2. REVISÃO DA LITERATURA

Nesta Seção, apresenta-se o modelo/teoria CST, enfatizando a definição e a exemplificação das relações, sobretudo as que codificam a complementaridade. Ademais, descrevem-se os principais trabalhos da literatura em que se propõem métodos para a identificação automática das relações CST ou semelhantes, destacando os atributos linguísticos nos quais eles se baseiam.

2.1. A teoria/modelo *Cross-document Structure Theory*

Inspirado na *Rhetorical Structure Theory*[4] (RST) (MANN; THOMPSON, 1987) e em um modelo teórico anterior (RADEV; MCKEOWN, 1998), Radev (2000) realizou uma análise de *corpus* para observar relacionamentos entre porções textuais de documentos que abordam mesmo assunto. Mais que propor novas relações a partir do modelo de Radev e McKeown (1998), Radev (2000) desenvolveu a CST.

A CST é um modelo semântico-discursivo multidocumento formado por um conjunto de relações que permite conectar (em pares) unidades informativas (p.ex.: sentenças) de textos distintos que abordam um mesmo assunto, explicitando, por exemplo, similaridades, complementaridades, contradições e variações de estilos de escrita entre as unidades dos pares.

Na proposta original, o modelo/teoria fornece um conjunto de 24 relações, as quais estão descritas no Quadro 3.

Quadro 3 - Conjunto original de relações CST.

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*

Fonte: Radev (2000).

[4] O objetivo da RST é analisar textos por meio da geração de árvores sintáticas com unidades de conteúdo (palavras, sentenças, ou mesmo parágrafos) que estejam relacionadas por alguma relação. Caso todas as unidades de conteúdo do texto estejam conectadas entre si, tem-se um texto coeso e coerente e, concomitantemente, com um nível informacional relevante.

Radev (2000) aponta que os relacionamentos entre documentos que abordam mesmo assunto podem ser estabelecidos em diversos níveis, a saber: lexical, sintagmático, sentencial e textual. Assim, as relações CST podem rotular conexões entre unidades informativas que pertencem a esses diferentes níveis. Em outras palavras, elas podem rotular conexões entre palavras, sintagmas, sentenças e documentos, e também entre parágrafos. Na Figura 3, ilustram-se os diferentes níveis em que as relações CST podem ser identificadas.

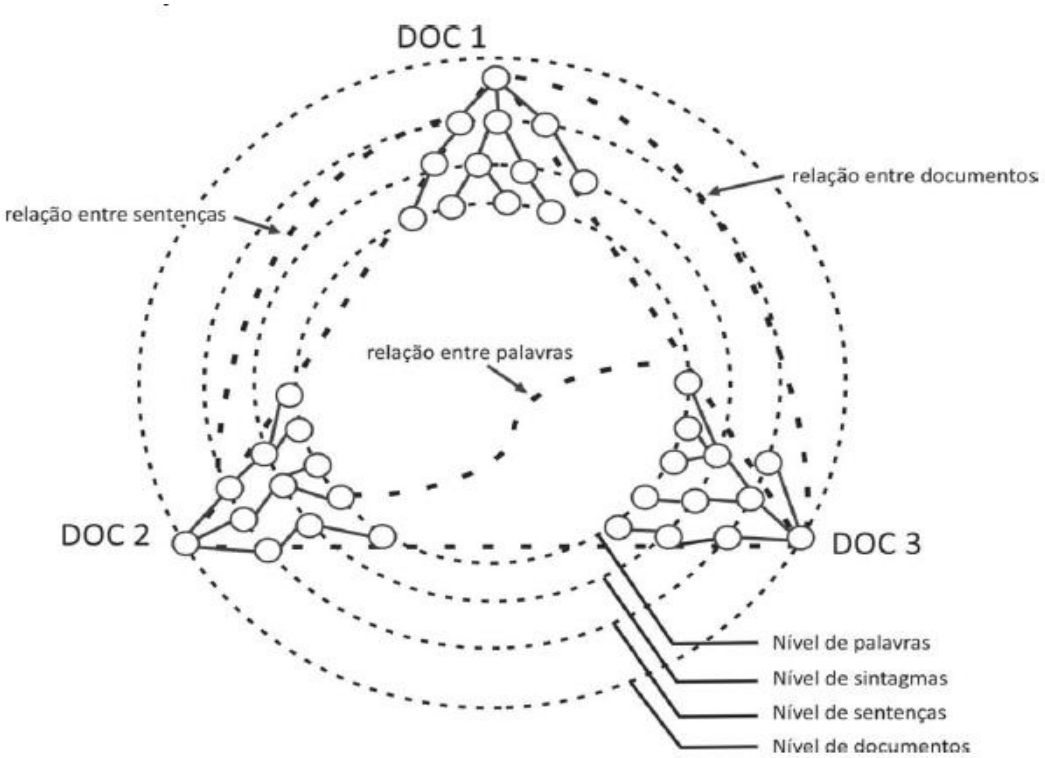

Figura 3 - Esquema de relacionamento CST.

Fonte: Radev (2000).

Especificamente, na Figura 3, vê-se que os níveis nos quais as relações CST podem ser identificadas compõem uma hierarquia (palavras → sintagma → sentença → texto), os quais estão representados por linhas pontilhadas. Assim, em cada nível da hierarquia, relações CST podem ser identificadas, ainda que usualmente isso seja feito em nível sentencial. Cada um dos 3 documentos (DOC 1, DOC 2 e DOC 3) está represento por um subgrafo, que codifica relações internas aos textos. Os relacionamentos internos a cada texto podem ser estabelecidos em nível sintático ou semântico-discurso (ou seja, por meio de uma teoria/modelo como a RST). As relações CST que podem ser

estabelecidas nos diferentes níveis, em especial, estão representadas por linhas pontilhadas mais grossas.

Sobre a CST, ressalta-se ainda que: (i) uma unidade de informação pode estar relacionada a várias outras unidades, ou seja, uma unidade pode apresentar mais de uma relação CST; (ii) nem todas as unidades textuais estão conectadas a outras, pois existem partes dos textos que não estão diretamente relacionadas a um mesmo tópico e, por isso, nem todas têm relações CST, (iii) os relacionamentos entre as unidades textuais podem ter direcionalidade e, consequentemente, as relações CST também podem.

No Quadro 4, há 2 trechos de textos, com 3 sentenças cada, provenientes de notícias jornalísticas distintas que relatam um mesmo acidente aéreo. Entre a sentença [1] do texto 1 e a sentença [1] do texto 2, identificam-se suas relações CST com direcionalidade.

Quadro 4 - Exemplos de relações CST.

Texto 1

[1] Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.

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

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

Texto 2

[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 3 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.

Fonte: http://www2.icmc.usp.br/~taspardo/sucinto/cstnews.html.

Por exemplo, a sentença [1] do texto 1 e a sentença [1] do texto 2 estão ligadas pela relação *Attribution*, pois tais sentenças apresentam informação em comum, sendo que a sentença [1] do texto 2 atribui essa informação a uma fonte/autoria (porta-voz das Nações Unidas). Outra relação entre as mesmas unidades também pode ser identificada.

No caso, a relação é a *Subsumption*, já que a sentença [1] do texto 2 apresenta, além do mesmo conteúdo da sentença [1] do Texto 1, informações adicionais.

Assim como na utilização de sua antecessora, a RST, a identificação de uma relação CST está sujeita a ambiguidades (AFANTENOS et al., 2004; ZHANG et al., 2002), pois, como toda análise subjetiva, pode haver mais de uma relação possível entre segmentos textuais. Com o objetivo de reduzir esta ambiguidade, alguns pesquisadores modificaram o conjunto original de Radev (2000).

Zhang et al. (2003) realizaram uma análise de *corpus* em inglês e, ao observarem a ambiguidade de algumas relações do conjunto original, propuseram a redução dos rótulos para 18, a saber: *Identity*, *Equivalence* (ou *Paraphrase*), *Translation*, *Subsumption*, *Contradiction*, *Historical Background*, *Citation*, *Modality*, *Attribution*, *Summary*, *Follow-up*, *Indirect speech*, *Elaboration (ou Refinement)*, *Fulfillment*, *Description*, *Reader profile*, *Change of perspective* e *Overlap* (ou *Partial equivalence*).

Aleixo e Pardo (2008), ao anotarem (em nível sentencial) um conjunto de textos jornalísticos em PB, eliminaram relações não verificadas no *corpus* e unificaram outras que foram consideradas muito similares, resultando em um conjunto de 14 rótulos (cf. Tabela 1, pág. 24). Como exemplo de unificação feita por Aleixo e Pardo (2008), as relações *Refinament*, *Description* e *Elaboration*, presentes no conjunto original, foram unificadas a um único rótulo genérico, *Elaboration.*

A partir do refinamento de Aleixo e Pardo (2008), Maziero et al. (2010) elaboram uma tipologia para as relações CST, ilustrada pela Figura 4.

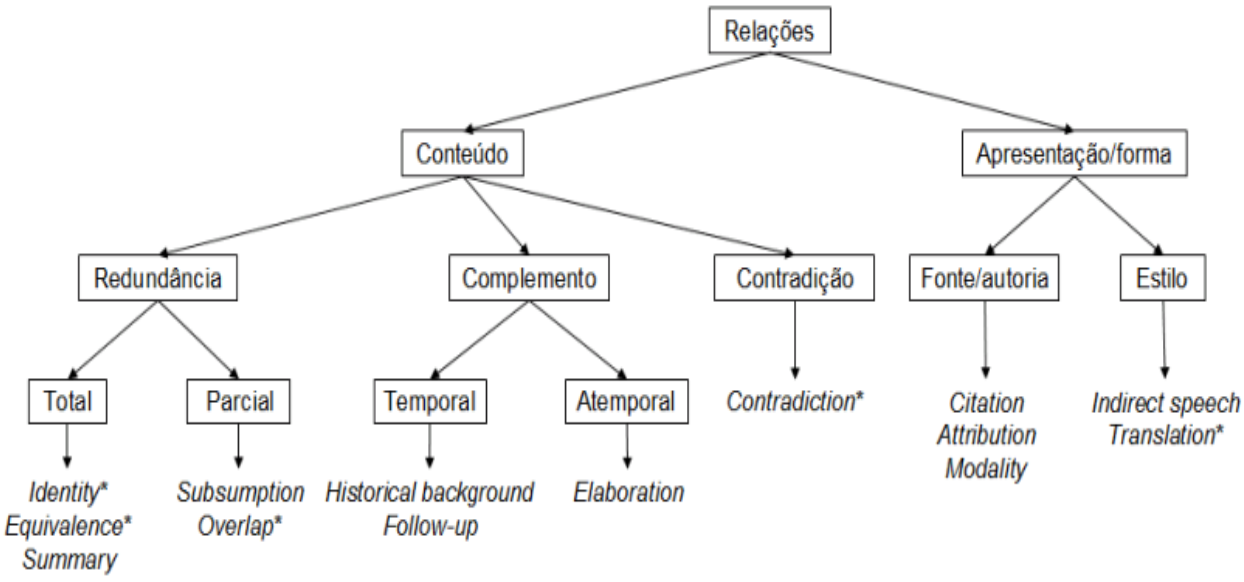

Figura 4- Tipologia das relações CST.

Fonte: Maziero et al. (2010).

Nessa tipologia, as relações CST foram organizadas em 2 grandes grupos: (i) relações de conteúdo (isto é, que rotulam os relacionamentos semânticos entre sentenças) e (ii) relações de forma (ou seja, que rotulam relacionamentos entre 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". Na Figura 4, o símbolo (*) indica que a relação não tem direcionalidade.

Com base nessa tipologia, vê-se, por exemplo, que as relações *Attribution* e *Subsumption* identificadas entre a sentença [1] do texto 1, e a sentença [1] do texto 2, presentes no Quadro 4 são, respectivamente, de forma e de conteúdo (em especial, de redundância parcial).

No Quadro 5, apresenta-se a definição de cada uma das 14 relações propostas por Maziero et al. (2010). Essa definição engloba 4 informações sobre a relação: (i) nome (ou rótulo), (ii) tipo, (iii) direcionalidade e (iv) restrição.

Quadro 5 - Definição das relações CST de Maziero et al. (2010).

Relação	**Tipo**	**Dir.**	**Restrições**	**Comentários**
Identity	Conteúdo→ Redundância Total	Nula	As sentenças devem ser idênticas	---
Equivalence	Conteúdo→ Redundância Total	Nula	As sentenças apresentam o mesmo conteúdo, mas expresso de forma diferente.	---
Summary	Conteúdo→ Redundância Total	S1 ← S2	S2 apresenta o mesmo conteúdo que S1, mas de forma mais compacta.	*Summary* é um tipo de *Equivalence*, mas *Summary* deve haver diferença significativa de tamanho entre as sentenças.
Subsumption	Conteúdo→ Redundância Parcial	S1 → S2	S1 apresenta as informações contidas em S2 e informações adicionais.	S1 contém X e Y, S2 contém X.
Overlap	Conteúdo→ Redundância Parcial	Nula	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.
Historical background	Conteúdo → Complemento Temporal	S1 ← S2	S2 apresenta informações históricas sobre algum elemento presente em S1.	O elemento explorado em S2 deve ser o foco de S2; se forem apresentadas informações repetidas, considere outra relação (p.ex.: *Overlap*); se os eventos em S1 e S2 forem relacionados, pondere sobre a relação *Follow-up.*

Follow-up	Conteúdo → Complemento Temporal	S1 ← S2	S2 apresenta acontecimentos que acontecem após os acontecimentos em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si.	---
Elaboration	Conteúdo → Complemento Atemporal	S1 ← S2	S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1.	O elemento elaborado em S2 deve ser o foco de S2; se forem apresentadas informações repetidas, considere outra relação (p.ex.: *Overlap*); se forem apresentadas informações temporais, pondere sobre a relação *Historical background*.
Contradiction	Conteúdo → Contradição	Nula	S1 e S2 divergem sobre algum elemento das sentenças.	---
Citation	Apresentação/ Forma → Fonte/Autoria	S1 ← S2	S2 cita explicitamente informação proveniente de S1.	Dada a natureza desta relação, ela não pode coocorrer com relações de redundância total.
Attribution	Apresentação/ Forma → Fonte/Autoria	S1 ← S2	S1 e S2 apresentam informação em comum e S2 atribui essa informação a uma fonte/autoria.	S1 e S2 apresentam informação em comum e S2 atribui essa informação a uma fonte/autoria.
Modality	Apresentação/ Forma → Fonte/Autoria	S1 ← S2	S1 e S2 apresentam informação em comum e em S2 a fonte/autoria da informação é indeterminada/relativizada/amenizada	Dada a natureza desta relação, ela não pode coocorrer com relações de redundância total.
Indirect speech	Apresentação/ Forma → Estilo	S1 ← S2	S1 e S2 apresentam informação em comum; S1 apresenta essa informação em discurso direto e S2 em discurso indireto.	---
Translation	Apresentação/ Forma → Estilo	Nula	S1 e S2 apresentam informação em comum em línguas diferentes.	---

Fonte: Maziero et al. (2010).

Como mencionado, o conjunto de 14 relações de Maziero et al (2010) foi proposto a partir da anotação manual de um conjunto de textos jornalísticos em português que gerou o *corpus* multidocumento denominado CSTNews (CARDOSO et al., 2011). Tal *corpus* será descrito em detalhes na subseção 3.1.1 (pág. 47).

Por ora, salienta-se que, no total, 1650 pares de relações CST foram manualmente identificadas no CSTNews, cuja distribuição é ilustrada na Figura 5. Destas, 1561 são da categoria de conteúdo. Se se somar a frequência das relações em

função das subcategorias de conteúdo (cf. Figura 4; pág. 21), tem-se a distribuição da redundância, complementaridade e contradição como ilustrada na Tabela 1.

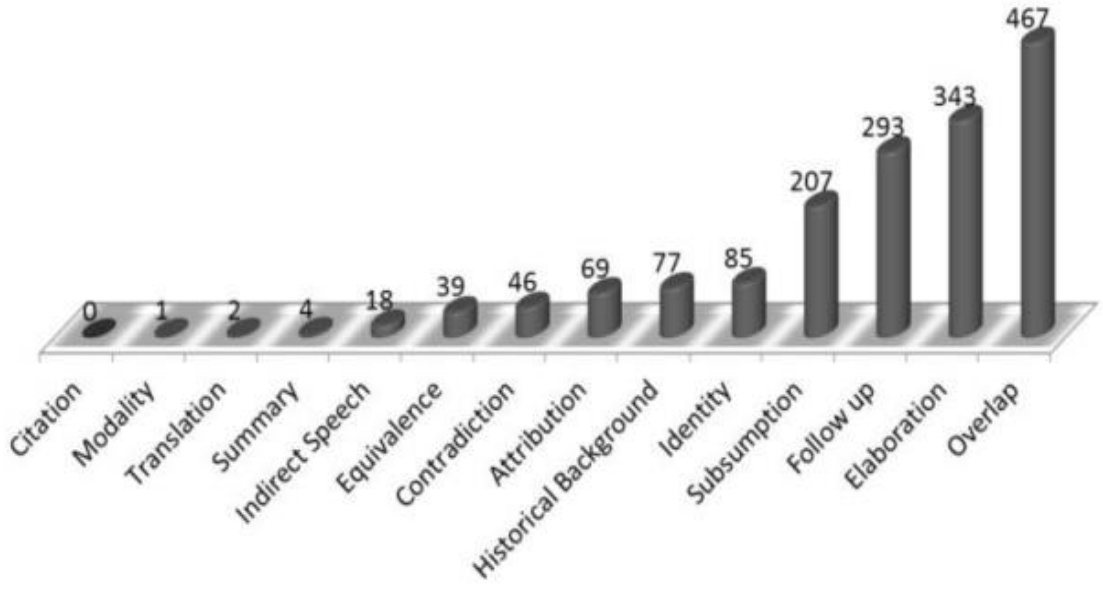

Figura 5- Frequência das relações CST no corpus CSTNews.

Fonte: Maziero (2012).

Tabela 1- Frequência das subcategorias de conteúdo no CSTNews.

Categoria	**Relação de conteúdo**	**Qt.**	**Total**
Redundância	*Identity*	85	802
	Equivalence	39	
	Summary	4	
	Subsumption	207	
	Overlap	467	
Compl.	*Follow up*	293	713
	Historical background	77	
	Elaboration	343	
Contradição	*Contradiction*	46	46

Fonte: Elaborado pelo autor.

Com base na Tabela 1, observa-se que, do total de 1650 pares de relações de conteúdo e de forma, 713 são da categoria complementaridade, o que equivale a 43%. Se se considerar apenas o total de relações de conteúdo (1561), a complementaridade representa 45,6% dos fenômenos multidocumento no *corpus.* Assim, vê-se que o fenômeno da complementaridade, observado por meio das relações CST, é bastante frequente em um *corpus* multidocumento. Isso ocorre porque esses fenômenos são identificados na relação entre textos que abordam mesmo assunto, nos quais a ocorrência de informações complementares é alta.

A seguir, descreve-se com mais detalhes a complementaridade com base na definição e exemplificação das relações CST dessa subcategoria.

2.2. As relações CST e a complementaridade

Como mencionado, a complementaridade denomina uma das subcategorias de conteúdo da tipologia de relações CST apresentada por Maziero et al. (2010). Os autores observaram que, na anotação do CSTNews, a complementaridade é o segundo fenômeno multidocumento mais frequente no *corpus,* com exceção da redundância.

De modo geral, compreende-se complementaridade pela relação que se estabelece entre duas sentenças, S1 e S2, sendo cada uma delas proveniente de um texto distinto, quando S2 apresenta informação complementar (ou seja, adicional ou suplementar) em relação a algum elemento presente em S1. Assim, na relação de complementaridade, uma das sentenças sempre possui informações adicionais em relação à outra. Em outras palavras, S1 e S2 podem possuir conteúdo em comum, sendo que S2 apresenta informação aditiva que não está presente em S1.

Ainda com base em Maziero et al. (2010), a complementaridade pode ser temporal ou atemporal.

As relações CST de complementaridade temporal podem ser de 2 tipos diferentes. Dado um par de sentenças, S1 e S2, as mesmas são complementares do subtipo temporal quando: (i) S2 apresenta informações históricas/ passadas sobre algum elemento presente em S1 e (ii) S2 apresenta acontecimentos/ eventos que sucederam os acontecimentos/ eventos presentes em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si. Os exemplos do

Quadro 6, retirados do *corpus* CSTNews, ilustram esses tipos de complementaridade.

Quadro 6 - Exemplos de complementaridade temporal.

Complementaridade temporal	Sentenças
(i) S2 apresenta informações históricas/ passadas sobre algum elemento presente em S1 (S1←S2)	S1: 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. S2: **Acidentes aéreos** são frequentes **no Congo**, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.
(ii) S2 apresenta acontecimentos/ eventos que sucederam os acontecimentos/ eventos presentes em S1 (S1←S2)	S1: **A pista auxiliar de Congonhas abriu** às 6h, apenas para decolagens. S2: **Congonhas só abriu** para pousos, às 8h50.

Fonte: Elaborado pelo autor.

A complementaridade do tipo (i) é ilustrada no Quadro 6 or um par de sentenças provenientes de textos que relatam “um acidente aéreo em Bukavu, no Congo”. Cada sentença é originária de um texto distinto da coleção. As sentenças do par estabelecem relação de complementaridade temporal porque S1 e S2 apresentam conteúdo comum (“acidente aéreo no Congo”), sendo que S2 apresenta uma informação adicional (histórica) sobre esse conteúdo que, nesse caso, diz respeito à “ocorrência frequente de acidentes aéreos no Congo (por causa do uso de aviões velhos)”. O conteúdo em comum entre as sentenças dos exemplos está negritado e o trecho de S2 que indica a informação suplementar está sublinhado. De acordo com a tipologia apresentada por Maziero et al. (2010), esse tipo de complementaridade temporal denomina-se *Historical background*.

A complementaridade temporal do tipo (ii) é ilustrada por um par de sentenças advindas de uma coleção cujos textos relatam “atrasos e cancelamentos no aeroporto de Congonhas devido ao mau tempo”. As sentenças estão em complementaridade temporal porque S1 e S2 apresentam informação comum (“abertura das pistas do aeroporto de

Congonhas" ou apenas "Congonhas"), sendo que S2 apresenta um acontecimento que sucedeu o evento descrito em S1 após um intervalo curto de tempo. No caso, S2 fornece "o horário de abertura da pista (principal) para pouso", que foi posterior ao evento de "abertura da pista auxiliar para decolagem" veiculado por S1. Segundo com a tipologia apresentada por Maziero et al. (2010), esse tipo de complementaridade temporal denomina-se *Follow-up*.

A relação de sequência temporal entre o evento focalizado em S2 e o evento descrito em S1 envolve a ocorrência de "expressões temporais" que, segundo Baptista et al. (2008), são do tipo "tempo_calendário" e subtipo "data" ("6h" e "8h50"). Tais expressões, no entanto, nem sempre ocorrem na complementaridade temporal, como pode ser visto no exemplo da relação de tipo (i) do Quadro 6.

As relações de complementaridade atemporal, ao contrário das exemplificadas no Quadro 6, não envolvem conteúdo que indica a localização no tempo (anterior ou posterior) de um acontecimento/fato em relação a outro. Essa complementaridade estabelece-se quando, dado um par de sentenças, S1 e S2, S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1. Além disso, o elemento elaborado em S2 deve ser o foco de S1. Os exemplos do Quadro 7, também retirados do *corpus* CSTNews, ilustram esse tipo de relação de conteúdo atemporal.

Quadro 7 - Exemplos de complementaridade atemporal.

Complementaridade temporal	**Sentenças**
S1 e S2, S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1 (S1 S2)	S1: Apesar da definição, o **cronograma da obra** não foi divulgado. S2: O **cronograma da obra** depende de estudos finais que estão sendo realizados pela Infraero.
	S1: As vítimas do acidente foram 14 passageiros e três **membros da tripulação**. S2: Segundo fontes aeroportuárias, os **membros da tripulação** eram de nacionalidade russa.

Fonte: Elaborado pelo autor.

No primeiro par de sentenças, sendo cada uma delas proveniente de um texto distinto que comunica a "reforma da pista principal do aeroporto de Congonhas", observa-se que S1 e S2 possuem conteúdo comum ("cronograma da obra"), sendo que S2 fornece uma informação adicional sobre esse conteúdo. A informação adicional em relação a S1 é o foco de S2 e consiste em "a razão pela qual o cronograma da obra não foi divulgado" ("dependente de estudos finais que estão sendo realizados pela Infraero").

No segundo par de sentenças, S1 e S2 também possuem conteúdo comum ("membros da tripulação"), sendo que S2 fornece uma informação adicional sobre os "membros da tripulação". A informação adicional em relação a S1, que é o foco de S2, diz respeito à "nacionalidade dos membros da tripulação", como pode ser visto no trecho sublinhado de S2 ("eram de nacionalidade russa") no Quadro 7. De acordo com Maziero et al. (2010), esse tipo de complementaridade atemporal denomina-se *Elaboration.*

Assim, observa-se que as informações adicionais dos exemplos são bastante variadas ("motivo/razão" e "nacionalidade"). No que se refere à realização linguística, a informação adicional em ambos os exemplos está expressa por meio de sintagmas verbais compostos por verbo ("depende" / "eram") e sintagma preposicional ("de estudos finais que estão sendo realizados pela Infraero" / "de nacionalidade russa").

Dada a relevância das relações CST, sobretudo na SAM, tem-se investigado a automatização do processo de identificação das mesmas, posto que a anotação manual é uma tarefa bastante custosa. Na próxima subseção, apresentam-se os principais trabalhos nos quais se propõem métodos para a detecção automática das relações CST, inclusive as de complementaridade, destacando as informações linguísticas utilizadas em tal tarefa.

2.3. Métodos de identificação automática das relações CST

Há vários trabalhos que propõem métodos para a identificação automática das relações semântico-discursivas da teoria/modelo CST ou de relações semelhantes. Dentre eles, destacam-se: Zhang et al. (2003), Zhang e Radev (2005), MacCartney et al. (2006), Miyabe et al. (2008) e Kumar et al. (2012) para o inglês; Marsi e Krahmer (2005), para o holandês; e Maziero (2012), para o PB. A seguir, tais trabalhos são descritos em detalhes, seguindo-se a ordem cronológica de publicação dos mesmos.

Nos métodos de Zhang et al. (2003) e Zhang e Radev (2005), a identificação das relações CST é feita em 2 etapas.

Baseando-se em Zhang e Radev (2005), observa-se que, na primeira, o método dos autores analisa se existe alguma conexão lexical entre as sentenças que compõem um par. Isso é feito porque já se observou que é improvável a ocorrência de relações CST entre sentenças que sejam lexicalmente muito diferentes. Para capturar a similaridade lexical (ou seja, o número de palavras em comum entre as sentenças), aplica-se a medida estatística *word overlap*, que é determinada pela aplicação da fórmula em (4). Caso o valor da *word overlap* obtido seja igual ou superior a 0.12[5], considera-se que as sentenças do par sob análise são relacionadas.

(4)

$$\text{WordOverlap}\,(S1, S2) = \frac{\#\text{Palavras em comum}}{\#\text{Palavras}(S1) + \#\text{Palavras}(S2)}$$

Em (4), 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*[6], números e símbolos). O resultado obtido será entre 0 e 0,5, sendo que, quanto mais próximo de 0,5 for a *Wol*, mais redundante será o par entre si, e, quanto mais próximo de 0, menos redundante.

Na segunda etapa, o método determina efetivamente a relação CST que ocorre entre as sentenças lexicalmente semelhantes que foram identificadas na etapa anterior. Para tanto, os autores se baseiam na similaridade de algumas características ou atributos entre as sentenças do par. Tais atributos são de diferentes níveis linguísticos. Especificamente, o método observa conjuntamente os seguintes atributos para determinar a relação CST: (i) número de palavras idênticas entre as sentenças (atributo lexical), (ii) número de classes de palavras idênticas (atributo sintático)[7], e (ii) distância semântica entre os núcleos de sintagmas nominais (SNs) e verbais (SVs) (atributo

[5] Com base em *corpus*, Zhang e Radev (2005) observaram que o valor de 0.12 para a medida *word overlap* era o "ponto de corte" (do inglês, *cutoff*) mais adequada para a detecção da similaridade.

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

[7] Essa similaridade é determinada pela quantidade de etiquetas morfossintáticas idênticas que há entre as sentenças de um par. As etiquetas morfossintáticas consistem em rótulos que indicam a classe das palavras (p.ex.: N(ome), ADJ(etivo), V(erbo), etc.), as quais são associadas às palavras de um texto de forma automática (isto é, *tagging*) ou manual.

semântico). Para determinar a distância semântica entre as palavras nucleares em SNs e SVs, o método utiliza a WordNet de Princeton[8] (WN.Pr), uma base relacional de dados lexicais (FELLBAUM, 1998).

No caso do atributo sintático, quanto maior o número de etiquetas em comum entre as sentenças, maior a similaridade entre elas. No caso do atributo semântico, a similaridade é determinada pela proximidade da relação que 2 núcleos de SNs, por exemplo, possuem na hierarquia de conceitos da WN.Pr. Por exemplo, caso 2 nomes estejam em relação direta de hiponímia, os SNs (e, consequentemente, as sentenças que os possuem) são considerados mais similares que os SNs cujos núcleos não estejam relacionados na WN.Pr ou estejam relacionados por conexões mais distantes.

Para avaliar o método, Zhang e Radev (2005) utilizaram um *corpus* de treinamento/teste composto por 6 coleções de textos, cujas principais características estão descritas no Quadro 8[9].

Quadro 8 - *Corpus* de treinamento e teste de Zhang e Radev (2005).

Coleção	**Tópico**	**Artigo**	**Tamanho (número de sentenças)**
Milan9	---	9	30
DUC	Biografia de John Lennon	4	46
Gulfair11	---	11	27
HKNews	Qualidade da água e ar	8	32
NIE	Armas nucleares da Coreia do Norte	5	14
Novelty	Câncer *and power lines*	4	21

Fonte: Zhang e Radev (2005).

As sentenças das referidas coleções foram manualmente anotadas com relações CST e os atributos necessários das sentenças para a determinação da similaridade (isto é, *word overlap*) e das relações CST (atributos lexical, sintático e semântico) foram explicitados. Além das 6 coleções do Quadro 8, os autores utilizam mais 1 coleção,

[8] A WN.Pr é uma base de dados lexicais em que as palavras e expressões do inglês americano estão organizadas em 4 classes: nome, verbo, adjetivo e advérbio. As unidades de cada classe estão codificadas em *synsets* (*synonym sets*), ou seja, conjuntos de formas sinônimas ou quase-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.

[9] No Quadro 8, ressalta-se que o nome dado às coleções reflete a fonte da qual os textos da coleção foram coletados

denominada *Shuttle10* (cujo tópico é o acidente o *Space Shuttle Columbia*, em 2003), cujas sentenças não foram anotadas via CST, mas os referidos atributos foram.

Na sequência, as 7 coleções do *corpus* foram submetidas a algoritmos de Aprendizado de Máquina (AM) que, a partir de atributos explícitos, aprende padrões estatisticamente relevantes e realiza os testes dos mesmos, os quais podem ser no próprio *corpus* de treinamento ou em outro *corpus* (de teste). No caso, as 7 coleções compuseram o *corpus* de treinamento e teste.

Os resultados dos testes realizados pelo AM são expressos pelas medidas clássicas de avaliação em PLN, a saber: precisão[10], cobertura[11] e medida-f[12] (HIRSCHMAN; MANI, 2003). No caso, o AM obteve os resultados descritos na Tabela 2, os quais incluem apenas as relações que tinham frequência maior que 20 nos dados de teste.

Tabela 2 - Resultados obtidos por Zhang e Radev na identificação automática de relações CST.

Relação CST	Precisão	Cobertura	Medida-F
No relation	0.8875	0.9605	0.9226
Equivalence	0.5000	0.3200	0.3902
Subsumption	0.1000	0.0417	0.0588
Follow-up	0.4727	0.2889	0.3586
Elaboration	0.3125	0.1282	0.1818
Description	0.3333	0.1071	0.1622
Overlap	0.5263	0.2941	0.3773

Fonte: Zhang e Radev (2005).

Na Tabela 2, observa-se que o reconhecimento de algumas relações é feito com precisão bastante baixa, como é o caso da relação *Subsumption*, cuja precisão é de 0.1. Segundo os autores, isso se deve à esparsidade dos dados de treinamento. Além disso, observa-se que, dentre as relações CST, estão 2 de complementaridade: *Follow up* (temporal) e *Elaboration* (atemporal). A precisão mais alta no reconhecimento automático da relação *Follow up* pode ser explicada pela natureza da própria relação, já que *Elaboration* é mais genérica que *Follow up* e, por isso, mais difícil de detectar.

[10] Precisão é o número de casos corretamente detectados em relação ao número total de casos detectados.
[11] Cobertura é o número de casos corretamente detectados em relação à quantidade que deveria ser detectada.
[12] Medida-f é a média ponderada dos cálculos de Precisão e Cobertura.

Marsi e Krahmer (2005), por sua vez, não focalizam a identificação específica de relações CST, mas de relações semelhantes: *Equals, Generalizes, Specifies, Restates* e *Intersects*, definidas no Quadro 9, com base no exemplo (5).

(5) Sentença 1: *Dailly coffe diminishes risk on Alzheimer and Dementia.*
Sentença 2: *Three cups coffee a day reduces chance on Parkinson and Dementia.*

Quadro 9 - Relações semânticas de Marsi e Krahmer (2005).

Relações	**Definição**	**Exemplo**
Equals	Sentenças (ou porções textuais) idênticas.	"*Dementia*" é idêntico a "*Dementia*"
Generalizes	O primeiro termo é mais geral que o segundo.	"*daily coffee*" é mais genérico que "*three cups of coffee a day*"
Specifies	O primeiro termo é mais específico que o segundo.	"*three cups of coffee a day*" é mais específico que "*daily coffee*"
Restates	Quando um elemento é paráfrase de outro.	"*risk*" é paráfrase de "*chance*"
Intersects	Quando, dado um par de sentenças, ambas compartilham alguma informação em comum, entretanto, algumas delas possui alguma informação não expressa no outro.	*"diminishes risk on Alzheimer and Dementia*" e "*reduces chance on Parkinson and Dementia*"

Fonte: Marsi e Krahmer (2005).

Os autores objetivam representar as sentenças de um par por meio de árvores de dependência e alinhar os nós das árvores que são semanticamente correspondentes por meio das relações do Quadro 9.

Para tanto, os autores representam os pares de sentenças por meio de árvores de dependência, a fim de alinhar os nós das árvores que são semanticamente correspondentes, e rotular as relações que se estabelecem entre os nós correspondentes utilizando as relações expressas no Quadro 9.

O *corpus* construído para a pesquisa é composto por duas traduções, em holandês, do livro "*Le petit prince*", escrito por Saint-Exupéry em 1943. De acordo com Marsi e Krahmer (2005), essa estratégia garante que o *corpus* tenha uma quantidade considerável de pares de sentenças relacionadas. Os cinco primeiros capítulos de cada

tradução foram definidos como *corpus* de treinamento, enquanto que o restante foi utilizado para teste. A parcela destinada a treinamento foi segmentada (a nível sentencial) e tokenizado. A partir disso, foram geradas árvores de dependência por meio de *parser*, e revisado manualmente.

O alinhamento foi realizado de forma semiautomática, baseando-se na similaridade das dependências (*head/subject*, *head/modifier* e *coordination/conjunction*). A similaridade foi identificada com base nas relações de sinonímia, hiperonímia e hiponímia da EuroWordNet[13] (VOSSEN, 1998).

O *corpus* de treinamento foi submetido a algoritmos de AM para a detecção das relações *Equals, Generalizes, Specifies, Restates e Intersects*. A fim de testar os padrões obtidos por AM, os autores automatizaram o processo de anotação, e o realizaram para a parcela do *corpus* destinada a teste. A partir desse processo, pôde-se inferir os resultados expressos na Tabela 3.

Tabela 3 - Avaliação estatística das relações propostas por Marsi e Krahmer (2005).

Relações	**Precisão**	**Cobertura**	**Medida-f**
Equals	0.93	0.95	0.94
Restates	0.56	0.78	0.65
Specifies	0.19	0.37	0.24
Generalizes	0.62	0.70	0.64
Intersects	n.a.	0	n.a.
Combinação	0.62	0.70	0.64

Fonte: Marsi e Krahmer (2005)

MacCartney et al. (2006) focam a identificação automática da relação de *entailment* (ou seja, o relacionamento estabelecido entre unidades de análise sob a forma de acarretamento) entre 2 sentenças, sendo que uma delas é denominada **hipótese**.

(6) Sentença 1: Estima-se que 2,5 a 3,5 milhões de pessoas morreram de AIDS no ano passado.

Sentença 2: Mais de 2 milhões de pessoas morreram de AIDS no ano passado

[13] A EuroWordNet é rede de relacionamento semântico de línguas europeias (Holandês, Italiano, Espanhol, Alemão, Francês, Checo e Estônio), inspirada na WordNet.Pr. Cada língua possui sua própria *wordnet*, entretanto estão todas interconectadas, a nível conceitual, pela interlíngua *Interligual Index*.

Em (6), a Sentença 1 expressa o texto, e a Sentença 2 a hipótese. Por meio de uma implicatura (ou acarretamento) semântica, a Sentença 2 está compreendida na Sentença já que, de fato, "mais de 2 milhões de pessoas" está compreendido por "2,5 a 3,5 milhões de pessoas". Assim, esse par de sentenças estabelecem-se como relacionado.

Para tanto, o método proposto pelos autores consiste em representar as sentenças de um par por meio de grafos de dependência, em que as palavras são codificadas pelos nós e as relações gramaticais estabelecidas entre elas são representadas por arestas. Na sequência, alinham-se os nós correspondentes das sentenças do par por meio de uma métrica que considera uma série de similaridades entre os nós, como (i) identidade dos lemas (ou canônica), (ii) identidade das classes de palavra e (iii) relações semânticas extraídas da WN.Pr. Uma vez que as sentenças tenham sido alinhadas, verifica-se se a hipótese é ou não acarretada pela sentença.

No método de MacCartney et al. (2006), o *entailment* é determinado por um conjunto de 28 características ou atributos linguísticos, os quais podem ser agrupados nas seguintes categorias:

a) Polaridade: marcadores linguísticos em contextos de polaridade negativa, expressos pela simples negação (por exemplo, "não"), quantificadores negativos ("menos"), preposições restritivas (como "exceto") e superlativos.
b) Adjunção (do inglês, *adjunct atributes*): marcadores que evidenciam o "abandono" ou a adição de adjuntos sintáticos, já que alguns adjuntos modificam "*Os cachorros latem*"[14] (em inglês, "*Dogs barked*") distingue-se de "*Os cachorros latem hoje*" (em inglês, "*Dogs barked today*") ou preservam ("*Os cachorros latem*" tem seu sentido compreendido em "*Os cachorros latem alto*") o sentido do texto para a hipótese.
c) Antonímia: marcadores que evidenciam a polaridade (negativa ou positiva) entre um par de antônimos advindos do texto e da hipótese. Para tanto, os autores propõem identificar os antônimos com base na WordNet.Pr e um alista de antônimos de referência. Observa-se, então, qual polaridade é expressa a partir do par de antônimos por meio do contexto do texto e/ou da hipótese.
d) Modalidade: marcadores que identificam a modalização entre o texto e a hipótese. Os autores analisam 6 modalizadores (a saber, *possible*, *not possible*, *actual*, *not*

[14] Tradução nossa.

actual, *necessary* e *not necessary*), e definem 5 julgamentos de relacionamento (a saber, *yes, weak yes, don't know, weak no* e *no*).

e) Factualidade: marcadores verbais que evidenciam pressuposições sobre um evento ("*O ladrão tentou escapar*" é diferente de "*O ladrão escapou*").

f) Quantificação (entre as sentenças): marcadores que evidenciam relação de quantificação entre o texto e a hipótese (*"Cada empresa deve informar a seus funcionários"; "Uma empresa deve informar a seus funcionários*"), as quais se dividiram em cinco categorias (a saber, *no, some, many, most* e *all*).

g) Tempo e data: marcadores que evidenciam a relação de tempo/data e entre o texto a hipótese (*Estima-se que 2,5 a 3,5 milhões de pessoas morreram de AIDS no ano passado.*).

h) Alinhamento: marcadores que identificam se o alinhamento de sentenças, entre o texto e a hipótese, está adequado. Para tanto, os autores propõem dois valores de qualidade ("*good score*" e "*bad score*"), os quais detectam se o alinhamento é "bom" ou "ruim", levando em consideração a distância entre o texto e a hipótese.

Para a avaliação dos atributos, MacCartney et al. (2006) utilizaram um conjunto de 567 pares de sentenças para treinamento, e outros 800 pares para teste. Por utilizarem uma representação de grafos, os autores mediram a precisão (acurácia) e a "pontuação ponderada" como métricas de avaliação dos atributos. Por meio dos atributos levantados, os autores geraram, então, grafos em que cada palavra de uma sentença é mapeada em pares de palavras de outra sentença, ou a nenhuma palavra. A acurácia máxima levantada por MacCartney et al. (2006) foi de 0,65.

Miyabe et al. (2008) também investigaram a identificação automática de relações semelhantes às do modelo CST. Em especial, os autores focaram nas relações *Equivalence* e *Transition*. Segundo os autores, *Equivalence* ocorre entre 2 sentenças quando estas veiculam a mesma informação por meio de palavras diferentes. *Transition*, por sua vez, ocorre entre 2 sentenças quando estas veiculam a mesma informação, mas apresentam distinção numérica.

Quadro 10 - Exemplo das relações Equivalence e Transition de Miyabe et al. (2008).

Texto 1
[1] ABC said on the 18th that the number of users of its mobile-phone service had reached 1.500,000.
[2] Users can acces the internet, reserve train tickets, as well as make phone calls through this service.
Texto 2
[1] ABC telephone company announced on the 9th that the number of users of its mobile-phone service had reached one million.
[2] This service includes internet access, and enables train-ticket reservations and telephone calls.

Fonte: Miyabe et al. (2008)

No

Quadro 10, de acordo com os autores, a primeira sentença de ambos os textos estabelece relação de *Transition* porque, apesar de transmitirem informação similar, o número de usuários expresso no Texto 2 varia em relação à quantidade expressa no Texto 1 ("*one milion*" e "*1.500,000*", respectivamente), já que a informação também está atribuída à variação de datas ("*on the 9th*" e "*on the 18th*"), respectivamente). Já a segunda sentença de ambos os textos estabelece relação de *Equivalence*, pois as sentenças transmitem a mesma informação, ainda que por meio de paráfrase.

Para identificar essas relações, os autores consideraram a similaridade entre as sentenças com base em: (i) quantidade de caracteres de cada sentença, (ii) data de publicação do texto-fonte de cada sentença, (iii) posição das sentenças nos texto-fonte, (iv) similaridade lexical (capturada pela medida do *cosseno*[15]), (v) similaridade semântica, (vi) conjunções, (vii) expressões ao final da sentença, (viii) entidade nomeada e (ix) tipo de entidade nomeada ("lugar", "hora", por exemplo).

Os autores utilizaram um *corpus* que compreende 115 conjuntos de textos jornalísticos inter-relacionados acerca de vários assuntos, organizados em 10 textos por coleção.

[15] A medida cosseno é resultado de uma representação de grafos de um texto, em que cada nó é uma sentença e as arestas são valores numéricos que apontam a proximidade entre duas sentenças, em relação ao léxico. Assim, quanto menor o ângulo entre duas sentenças, há maior similaridade entre elas.

Para identificação da relação *Elaboration*, os autores anotaram o corpus utilizando o modelo CST, e observaram que, de um pouco mais que 470.000 pares de sentenças, 798 possuíam tal relação.

Para identificação da relação *Transition*, os autores propuseram um algoritmo[16], a saber: (i) identificar os sintagmas nominais (SN) constituídos por valores numéricos, (ii) identificar as frases em que os valores números são dependentes em sintagmas preposicionais (SP), (iii) buscar os SNs que dependem dos SPs e (iv) extrair os SNs encontrados em (iii), exceto informações sobre data. No

Quadro 10, "*one milion*" e "*1.500,000*" são valores numéricos, e "*had reached*" é o SP. Assim "*the number of users of its mobile-phone service*" seria o SN em que os valores numéricos dependem.

A avaliação foi realizada com base nas medidas precisão (P), cobertura (C) e medida-f (MF). Para a identificação da relação *Elaboration*, o método obteve P = 87,2; C = 57,3; MF = 69,2, enquanto que, para a detecção de *Transition*, o método obteve P = 27,4; C = 41,2; MF = 32,9.

Dentre os trabalhos que se detiveram a identificar automaticamente as relações CST ou semelhantes, destaca-se o de Maziero (2012), do qual resultou o CSTParser, um analisador discursivo para textos em PB. Nessa ferramenta de PLN, as relações CST são identificadas com base nos atributos linguísticos até então mais difundidos da literatura. Especificamente, o método de análise multidocumento de Mazeiro (2012) identifica as relações CST com base nas informações descritas no Quadro 11.

Quadro 11 - Atributos de Maziero (2012).

1	Diferença de tamanho em palavras (S1-S2)
2	Porcentagem de palavras em comum em S1
3	Porcentagem de palavras em comum em S2
4	Posição de S1 no texto (0- início, 2- fim, 1- meio)
5	Número de palavras na maior *substring* entre S1 e S2
6	Diferença no número de substantivos entre S1 e S2
7	Diferença no número de advérbios entre S1 e S2
8	Diferença no número de adjetivos entre S1 e S2
9	Diferença no número de verbos entre S1 e S2
10	Diferença no número de nomes próprios entre S1 e S2
11	Diferença no número de numerais entre S1 e S2

[16] Algoritmo, de maneira geral, trata-se de uma sequência de passos predeterminados para realizar uma tarefa.

12	Sobreposição de sinônimos entre S1 e S2

Fonte: Maziero (2012).

Vale ressaltar que, além das características sentenciais do Quadro 11, o método de Maziero (2012) utiliza regras específicas para a identificação das relações *Identity, Contradiction* (explícita), *Attribution, Indirect Speech* e *Translation*. Para ilustração, destaca-se que a regra formulada para a identificação da relação *Contradiction* prevê apenas os casos de contradição do tipo explícita, a saber, resultantes de diferenças numéricas entre as sentenças de um par. Por exemplo, caso haja um símbolo do tipo hora ("h") (ou medidas como metros, quilômetros, etc.) nas sentenças de um par, verifica-se se os valores vinculados a esses símbolos são iguais ou diferentes. Se diferentes, a regra indica que há uma contradição entre as sentenças.

Para avaliar o método, o autor utilizou o *corpus* CSTNews (CARDOSO et al., 2011). Em linhas gerais, o CSTNews é um *corpus* multidocumento de textos jornalísticos em PB. Tal conjunto de textos está organizado com base na teoria CST, em que cada *cluster* (isto é, conjunto de textos separados por assunto) possui, em média, 3 textos que abordam o mesmo tema.

De acordo com Mazeiro (2012), foi realizado um treinamento num período de 3 meses, em que ao final de cada treinamento as relações CST obtidas por cada autor eram discutidas pelo grupo, a fim de avaliar a concordância e a compreensão da tarefa. Após isso, os anotadores se organizaram em três grupos e anotaram a parcela restante do *corpus*, em que cada anotador realizava a anotação em cada um dos três textos do *cluster*.

Ao avaliar o desempenho do método, Mazeiro (2012) obteve a precisão de geral de 68,13%. Essa precisão geral é a média ponderada da precisão dos atributos do Quadro 11 para a identificação das relações *Overlap, Subsumption, Elaboration, Equivalence, Historical background* e *Follow up* (de conteúdo) e da precisão das regras para a identificação das relações *Identity, Contradiction* (explícita), *Attribution, Indirect Speech* e *Translation*[17]. Segundo o autor, essa precisão é considerada boa devido à subjetividade inerente à tarefa de identificação das relações multidocumento.

Ainda segundo o autor, as relações *Follow-up* e *Equivalence* são classificadas equivocadamente como *Overlap*, já que o grau de similaridade de elementos na

[17] Ressalta-se que as relações *Summary, Modality* e *Citation* não foram consideras no método de Maziero (2012) devido à baixa frequência no *corpus* utilizado, o CSTNews.

superfície textual pode ser bastante semelhante. A relação *Historical background* pode ser confundida com a relação *Elaboration*, pois ambas podem ter informações temporais. O autor ainda aponta que esses equívocos ocorrem por conta da falta de atributos que descrevam tais relações de forma específica e possam distinguir com mais exatidão uma relação da outra.

Nos trabalhos de Kumar et al. (2012), tem-se um método para a identificação de somente 4 relações CST provenientes do conjunto original: *Identity*, *Overlap*, *Subsumption* e *Description*. Considerando-se a tipologia de Maziero et al (2010), essas relações são da categoria de conteúdo, uma vez que *Description*[18] (juntamente com *Refinament*) foi fundida à relação *Elaboration*.

O método de Kumar et al. (2012) pauta-se em 4 características sentenciais: (i) similaridade lexical, capturada pelas medidas distintas *cosseno* e *word overlap;* (i) tamanho das sentenças; (iii) similaridade de sintagma nominal, e (iv) similaridade de sintagma verbal. Para avaliar o método, os autores utilizam 476 pares de sentenças para treinamento e 206 pares para teste, todos provenientes do CSTBank[19] (RADEV, 2003). O conjunto de teste inclui 100 pares compostos por sentenças sem anotação de relações CST.

A partir da explicitação das 4 características (ou atributos) relativas às sentenças do *corpus* de treinamento, 3 algoritmos distintos de AM foram utilizados para o aprendizado de padrões estatisticamente relevantes de detecção das relações. Tais padrões foram aplicados ao conjunto de teste e os resultados obtidos pelos 3 algoritmos revelam, de modo geral, boa performance na identificação da relação *Identity* (i.e. medida-f > 90%) e na detecção dos pares sem relação CST (isto é, medida-f > 80%).

Segundo os autores, esses resultados podem ser decorrentes do fato de as sentenças relacionadas por *Identity* apresentarem alta similaridade lexical e de tamanho e as sentenças sem relação CST não apresentarem tais características. Na verdade, as sentenças sem relação CST apresentam características opostas.

Da revisão sobre os trabalhos em que foram propostos métodos automáticos de identificação das relações multidocumento CST ou semelhantes, observa-se que os métodos pautam-se fortemente na similaridade ou redundância entre as sentenças do

[18] A relação *Description* é descrita da seguinte forma: "S1 descreve uma entidade mencionada em S2" (KUMAR et al., 2012).

[19] *Corpus* multidocumento composto por textos jornalísticos em inglês cujas sentenças foram manualmente anotadas com as relações CST.

par. Isso se deve ao fato de as relações do tipo CST, sobretudo as de conteúdo, estabelecerem-se entre sentenças que de fato possuem sobreposição de conteúdo em diferentes graus ou níveis.

Assim, na sequência, apresentam-se os principais trabalhos em que métodos automáticos de identificação da redundância ou similaridade entre sentenças são propostos, com ênfase na descrição das características ou atributos linguísticos por eles empregados.

2.4. Métodos de identificação automática da similaridade

Quanto à detecção da redundância, ressalta-se que há vários trabalhos que descrevem diferentes métodos, como os de Hatzivassiloglou *et al.* (1999, 2001), Newman *et al*. (2004) e Hendrickx *et al*. (2009), para o inglês, e Souza et al (2012), para o português.

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* (cf. (4); pág. 29).

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:

a) sobreposição de etiquetas morfossintáticas: identifica etiquetas morfossintáticas em comum, sejam elas rótulos para as palavras de classe aberta (p. ex.: N, ADJ, etc.) como para as palavras de classe fechada (p.ex.: CONJ(unção), PREP(osição), etc.).
b) 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 *stem overlap*.

c) 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 *noun phrase head overlap*.

d) sobreposição de palavras sinônimas: identifica a similaridade em função da sobreposição de palavras semanticamente relacionadas (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")[20]. Tendo em vista que a identificação da sobreposição de palavras sinônimas para o inglês é feita com base na WN.Pr, a medida é especificada como *WordNet overlap*.

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:

a) sobreposição de palavras + ordem: busca-se verificar se as palavras em comum em uma sentença ocorrem na mesma ordem na outra sentença do par.

b) 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, objetiva-se identificar sobreposição de colocações. Caso a janela tenha tamanho 5, por exemplo, identificam-se palavras relacionadas em uma região da sentença.

[20] 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/.

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 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. 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, no qual a redundância é calculada pela similaridade semântica entre palavras alinhadas em nível sintático. Para tanto, os autores partem de um *corpus* comparável monolíngue[21] 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 sinonímia 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.

[21] Um *corpus* comparável monolíngue é composto por dois ou mais *subcorpora* com textos originais em suas respectivas línguas.

Para o português, Souza et al (2012) analisaram 3 grupos de atributos simples[22] (i) superficial estatístico, (ii) superficial linguístico e (iii) estrutural, totalizando 9.

Os 3 atributos superficiais estatísticos investigados pelos autores foram: (i) sobreposição de palavras, (ii) sobreposição de nomes e (iii) sobreposição de verbos.

Para o cálculo do atributo (i), utilizou-se a medida *word overlap* (Wol), obtida pela fórmula apresentada em (4) (pág. 29). Os demais atributos foram calculados com base em variações da *word overlap* em função das classes de palavras "nome" e "verbo". Para o cálculo da *word overlap* em função da classe de palavra "nome", por exemplo, a fórmula em (4) foi adaptada, gerando-se a fórmula *noun overlap* (Nol) descrita em (7). De forma análoga, fez-se a adaptação da medida original para o cálculo da sobreposição dos verbos, originando a fórmula *verb overlap* (Vol).

(7)

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

Os 5 atributos superficiais linguísticos, por sua vez, foram: (i) sobreposição de padrões morfossintáticos (PdMorf), (ii) sobreposição de verbo principal (Vp), (iii) sobreposição de núcleo de sujeito (Suj), (iv) sobreposição de núcleo de objeto/predicativo principal (ObjPredp), e (v) sobreposição de etiquetas morfossintáticas (EtMorf).

O atributo "sobreposição de padrões morfossintáticos" busca identificar a ocorrência em comum nas sentenças de unidades lexicais complexas e colocações. Em outras palavras, buscou-se identificar padrões morfossintáticos como [N_ADJ_PREP_N], [N_PREP_N_ADJ], [N_PREP_N] e [N_ADJ], etc.

O atributo "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 atributo.

Os atributos sobreposição de "núcleo de sujeito" e "núcleo de objeto/predicativo principal" também não foram explicitamente citados na literatura. No entanto, tais atributos são vistos como uma especificação do atributo "sobreposição de núcleo de

[22] Os autores evitaram a utilização de atributos cuja identificação necessitasse de tarefas complexas de pré-processamento de *corpus*, como o alinhamento de árvores sintáticas.

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 atributo "sobreposição de palavras sinônimas" foi selecionado por ser um dos mais utilizados na literatura para capturar a similaridade entre sentenças no nível semântico. Para tanto, várias fontes de conhecimento lexical do português, digitais e impressas, foram utilizadas, 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 o (iii) *Dicionário Eletrônico Houaiss da Língua Portuguesa* (HOUAISS, VILLAR, 2001).

Por fim, o atributo "sobreposição de etiquetas morfossintáticas" busca capturar a similaridade entre as sentenças com base no nível morfossintático, sem considerar a ordem de ocorrência das etiquetas, diferenciando-se, assim, do atributo "sobreposição de padrões morfossintáticos".

Aos atributos estatísticos e linguísticos, acrescentou-se outro, classificado como "superficial estrutural", a saber: "sobreposição de localização" (Loc). Este foi proposto 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. Consequentemente, quanto mais próximas forem as posições das sentenças em seus respectivos textos-fonte, maior a chance de serem similares. Essa similaridade foi 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.

Souza et al. (2012) testaram os 9 atributos em um conjunto composto por 45 pares de sentenças extraídas do *corpus* CSTNews. Esse conjunto de sentenças se distribui como descrito na Tabela 4.

Tabela 4 - Características do corpus de treinamento e teste de Souza et al. (2012).

Tipo de relação	**Relação**	**Quantidade de pares por relação**	**Quantidade de pares por nível de redundância**
Redundância total	*Identity*	5	15
	Equivalence	6	
	Summary	4	
Redundância	*Subsumption*	8	16

parcial	*Overlap*	8	
Não-redundância	----	14	14

Fonte: Souza et al. (2012).

Os 9 atributos de cada uma das sentenças foram manualmente descritos e, na sequência, calculou-se a similaridade entre as sentenças com base nos atributos explicitados.

Na sequência, os resultados do cálculo da similaridade foram submetidos ao ambiente de AM denominado *Waikato Environment for Knowledge Analysis* (Weka) (FRANK et al., 2011) com o objetivo de investigar a adequação dos atributos s quanto à identificação de: (i) os níveis de redundância (total, parcial e nula) e (ii) as relações CST de redundância.

Dentre os algoritmos do Weka, salienta-se que os testes foram realizados com o algoritmo PART, que gera regras no formato lógico *se, então*. Especificamente, realizaram-se 11 testes. No Teste 1, os 9 atributos foram submetidos em conjunto. Nos Testes 2, 3, 4, 5, 6, 7, 8, 9 e 10, os atributos foram testados individualmente. No Teste 11, os atributos de melhor desempenho individual (Wol e Nol) foram submetidos em conjunto.

Na Tabela 5, apresentam-se os testes ranqueados em função da precisão obtida para a distinção dos níveis de redundância (total, parcial ou nula).

Tabela 5 - Resultados depreendidos por meio de AM de Souza et al. (2012).

Teste	Atributo									Precisão (%)
	Loc	Wol	Nol	Vol	PdMorf	Suj	Vp	ObjPredp	EtMorf	
1	■	■	■	■	■	■	■	■	■	97,7
4			■							91,1
11		■	■							91,1
3		■								80
8							■			57,7
5				■						55,5
10									■	55,5
6					■					53,3
7						■				53,3
9								■		46,6
2	■									42,2

Fonte: Souza et al. (2012).

No Teste 1, a aplicação conjunta dos 9 atributos gerou as regras em (8), que obtiveram 97,7% de precisão. Dos 45 pares, apenas 1 foi classificado erroneamente pela aplicação da regra 4.

(8)

1. Se Nol<=0.09 então nulo (14 acertos)
2. Senão se Vp = não e EtMorf <=0 .9 e Loc <=0.27 então parcial (12 acertos)
3. Senão se Vp=sim então total (11 acertos)
4. Senão se PdMorf<=0.33 então total (5 acertos /**1 erro**)
5. Senão parcial (3 acertos)

Observa-se que as regras pautam-se em 5 atributos específicos para discriminar os tipos de redundância, a saber: *noun overlap*, *verb overlap*, sobreposição de localização e sobreposição de padrões morfossintáticos e sobreposição de etiquetas morfossintáticas com 97,7% de precisão. As regras geradas somente com base nos valores de Nol obtiveram os mais altos índices de precisão individual, 91,1% (Teste 4).

Quanto às relações CST, o algoritmo PART gerou um conjunto de regras pautado nos mesmos 5 atributos para identificar as relações de redundância (*Identity, Equivalence, Summary, Subsumption* e *Overlap*) com precisão de 88,88%.

Do que foi exposto, observa-se que as relações CST são identificadas, na maioria dos trabalhos, em função de alguns atributos linguísticos que capturam apenas a similaridade entre 2 sentenças de um par. Isso se justifica porque tais relações efetivamente estabelecem-se entre sentenças que apresentam certa sobreposição de conteúdo.

Por outro lado, certas relações de conteúdo, como *Contradiction* e *Identity* no trabalho de Maziero (2012), por exemplo, são identificadas por meio de regras que capturam as características específicas dessas relações.

Mesmo a complementaridade sendo um fenômeno frequente em um *corpus* multidocumento e cujo tratamento pode subsidiar a produção de sumários automáticos informativos e coerentes, este ainda não foi analisado a ponto de se identificar características mais refinadas que possam gerar regras automáticas de detecção do mesmo. Até o momento, o que se sabe sobre a complementaridade diz respeito à própria definição das relações (cf. RADEV, 2000; MAZEIRO, 2012).

Assim sendo, propõe-se o trabalho descrito a seguir. Na Seção 3, apresentam-se os objetivos da pesquisa e as suas hipóteses, além de descreverem-se as etapas metodológicas realizadas, as em andamento e as futuras.

3. PROPOSTA

Diante do cenário apresentado na revisão da literatura, propõe-se descrever as características específicas da complementaridade e propor métodos automáticos de detecção da complementaridade com base nessas características. Esses objetivos foram formulados porque se acredita na hipótese principal de que eles possam, juntamente com a redundância, subsidiar métodos automáticos de detecção dos diferentes tipos de complemento (temporais e atemporais) e das relações CST que expressam complemento (*Historical background*, *Follow-up* e *Elaboration*).

Para tanto, estabeleceram-se as seguintes etapas metodológicas: (i) seleção e recorte do *corpus;* (ii) análise do fenômeno da complementaridade em *corpus* para a delimitação de atributos linguísticos que o caracterizem*;* (iii) caracterização do *corpus*, que consiste na descrição/ explicitação dos atributos que tipificam a complementaridade; (iv) identificação ou proposição de atributos de detecção da complementaridade (e das respectivas relações CST); (v) estudo da correlação entre os atributos e os tipos de complementaridade; (vi) estudo da correlação entre os atributos e as relações CST, e (vii) avaliação dos atributos de melhor desempenho identificados em (v) e (vi).

Destas, as tarefas (i), (ii), (iii) e (iv) foram realizadas para uma parcela do *corpus* selecionado, as quais são descritas na sequência. Ressalta-se que, antes da atividade (iv), que consiste na proposição de métodos de detecção automática da complementaridade, uma análise manual da pertinência dos atributos delimitados foi realizada a partir da caracterização de uma parcela do *subcorpus.*

3.1. Tarefas realizadas

3.1.1. Seleção do corpus e construção do subcorpus

Para a realização da pesquisa, selecionou-se o CSTNews (CARDOSO et al. 2011), único *corpus* multidocumento de textos jornalísticos em PB anotado com as relações do modelo CST.

O CSTNews está organizado em 50 *clusters* (ou coleções) de textos jornalísticos (notícias) que abordam mesmo assunto, sendo cada um deles proveniente de uma fonte de notícias distinta. No total, o CSTNeews possui 140 textos, que somam 2.088 sentenças e 47.240 palavras. Os textos foram coletados dos seguintes jornais *online*:

Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e *Gazeta do Povo*. Essas fontes foram escolhidas devido à popularidade e circulação na *web*, garantindo a coleta de uma mesma notícia veiculada por fontes distintas.

Os *clusters* no CSTNews estão organizados em categorias, cujos rótulos indicam a seção do jornal da qual os textos que os constituem foram compilados. Assim, têm-se as categorias "mundo", "política", "cotidiano", "ciência" e "esporte".

Cada *cluster*, em especial, possui 2 ou 3 textos-fonte, sumários monodocumento e multidocumento de referência (manuais) e automáticos, alinhamento manual das sentenças dos sumários multidocumento às suas respectivas sentenças dos textos-fonte e uma série de anotações linguísticas. Dentre elas, estão: (i) relacionamentos semântico-discursivos multidocumento para cada conjunto de textos via relações CST; (ii) anotação de expressões temporais dos textos-fonte; (iii) etiquetação morfossintática (ou *tagging*); (iv) anotação dos sentidos dos substantivos e verbos de cada *cluster*; (v) anotação de aspectos informacionais nos sumários multidocumento (*o quê*, *onde*, *quando*, por exemplo), (vi) anotação semiautomática dos textos-fonte pela teoria RST e (vii) anotação manual de subtópicos informativos em cada texto-fonte do corpus.

Todos os sumários do CSTNews, mono e multidocumento, possuem taxa de compressão de 70% (ou seja, apresentam 30% de seus respectivos textos-fonte). No caso dos sumários multidocumento, a taxa de compressão de 70% é calculada com base no maior texto do *cluster*, medido em número de palavras.

No Quadro 12, tem-se a distribuição dos *clusters* em função de sua categoria.

Quadro 12 - Distribuição dos clusters nas categorias do CSTNews.

Categoria/Assunto	***Cluster* (C)**
Mundo	C1, C10, C12, C13, C14, C15, C18, C23, C26, C29, C32, C35, C46, C47
Política	C2, C9, C16, C17, C20, C21, C40, C42, C43, C44, C50
Cotidiano	C3, C4, C5, C6, C11, C22, C33, C34, C36, C37, C39, C45, C49
Ciência	C7
Esportes	C8, C19, C24, C25, C27, C28, C31, C38, C41, C48

Fonte: Cardoso et al. (2011).

Quanto à anotação CST, ressalta-se que esta foi realizada por 4 anotadores (linguistas computacionais) durante 3 meses. Para tanto, construiu-se a ferramenta CSTTool que, dado um *cluster*: (i) segmenta os textos-fonte em nível sentencial, (ii) identifica, em

pares, sentenças lexicalmente relacionadas por meio da medida *word overlap*, e (iii) disponibiliza ao anotador um conjunto de 14 relações CST (ALEIXO; PARDO, 2008). Para tais relações, Mazeiro et al. (2010) propuseram a tipologia da Figura 4 (pág. 21).

No Quadro 13, mostra-se a porcentagem de cada relação no CSTNews. Vale ressaltar que algumas relações ocorrem com frequência baixa ou não ocorrem no *corpus* devido à tipologia do mesmo. A relação *Translation*, por exemplo, não ocorre no CTSNews (frequência 0%), posto que se trata de um *corpus* monolíngue.

Quadro 13 - Frequência de ocorrência das relações no CSTNews.

Relação	**Frequência no *corpus***
Elaboration	23,98%
Overlap	19,85%
Subsumption	15,24%
Background	6,49%
Atributtion	5,68%
Equivalence	5,09%
Follow-up	4,72%
Contradiction	4,35%
Summary	4,35%
Identity	3,69%
Modality	3,54%
Indirect Speech	2,73%
Citation	0,29%
Translation	0%

Fonte: Cardoso et al. (2011).

Para estudar especificamente a complementaridade, fez-se um recorte no CSTNews, o qual consistiu em selecionar apenas os pares de sentenças anotadas com as relações CST do tipo complementaridade, ou seja, *Follow-Up*, *Historical Background* e *Elaboration*. Esse recorte foi feito por meio da interface *online* de consulta ao *corpus*[23].

Desse recorte, obteve-se, em termos absolutos, um conjunto de pares de sentenças com a seguinte distribuição: (i) 343 pares de complementaridade atemporal, ou seja, 343

[23]Disponível em: http://nilc.icmc.usp.br/CSTNews/

pares de sentenças relacionadas via *Elabaration*, e (ii) 370 pares de complementaridade temporal, sendo 293 pares de sentenças anotados com a relação *Follow-Up* e 77 pares cujas sentenças estão em relação de *Historical background.*

No Quadro 14, apresentam-se os dados quantitativos do *subcorpus.*

Quadro 14 - Dados quantitativos do subcorpus de análise.

Complementaridade	**Relação CST**	**Qt. de pares**	**Total**
Atemporal	*Elaboration*	343	343
Temporal	*Follow-up*	293	370
	Historical background	77	
--	--	--	**713**

Fonte: Elaborado pelo autor.

Na Figura 6, apresenta-se a distribuição percentual dos pares do *subcorpus* em função das relações CST.

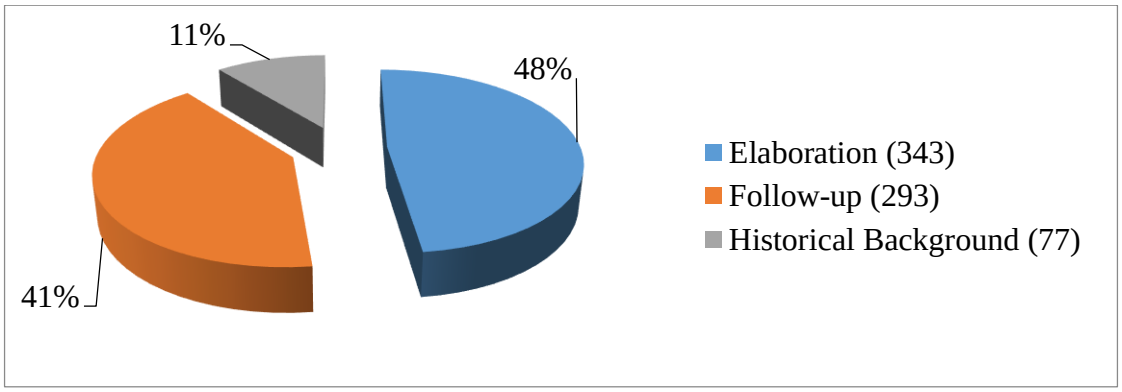

Figura 6 - Distribuição das relações de complementaridade no subcorpus em termos percentuais.

Fonte: Elaborado pelo autor.

Tendo em vista os 713 casos de complementaridade do *subcorpus*, realizou-se uma análise preliminar dos casos. Especificamente, fez-se uma análise manual de 135 pares de sentenças, sendo: (i) 45 pares cujas sentenças estão anotadas com a relação *Elaboration* (isto é, complementaridade atemporal), (ii) 45 pares com relação *Historical background* e (iii) 45 com relação *Follow up* (isto é, complementaridade temporal). Frente aos mais de 700 casos de complementaridade, 135 parece não ser uma quantidade expressiva de casos para caracterizar o fenômeno em questão. Entretanto, observou-se que a complementaridade, sobretudo a do subtipo temporal, manifesta-se

de forma regular no *subcorpus* e, por isso, a análise manual de um conjunto relativamente pequeno de casos pode revelar características típicas do fenômeno.

Na próxima subseção, descreve-se a análise manual dos 135 casos do *subcorpus*.

3.1.2. Análise preliminar da complementaridade em corpus

3.1.2.1. Características gerais da complementaridade

Maziero et al. (2010) apontam que a complementaridade pode envolver aspectos temporais e atemporais.

A complementaridade temporal, em especial, envolve certa sobreposição de conteúdo entre as sentenças de um par, sendo que uma das sentenças apresenta informação adicional com relação ao sobreposto. Esse acréscimo pode consistir de um acontecimento subsequente ao evento principal descrito em uma das sentenças. Essas informações adicionais se manifestam, segundo a tipologia de Mazieto et al. (2010), por meio de evidências linguísticas que evocam aspectos temporais.

A complementaridade atemporal também se caracteriza por certa sobreposição de conteúdo entre as sentenças de um par e por uma das sentenças fornecer informação adicional sobre o tópico principal. No entanto, o que a diferencia da complementaridade temporal é o fato de que a informação adicional nem sempre é marcada por uma evidência linguística na superfície textual, especialmente relativa a aspectos temporais, como aponta Maziero *et al.* (2010).

Assim, com base na literatura pesquisada e na análise manual da parcela inicial do *corpus* composta por 135 pares de sentenças, identificam-se 3 atributos potencialmente relevantes para a caraterização geral do fenômeno da complementaridade, a saber: (i) a redundância entre as sentenças, (ii) a localização das sentenças no texto-fonte e (iii) o subtópico veiculado pelas sentenças.

a) <u>A redundância ou similaridade lexical</u>

Sabe-se que as relações CST sempre ocorrem entre sentenças que são semanticamente relacionadas. Em outras palavras, elas ocorrem entre sentenças que possuem conteúdo em comum, em menor ou maior grau, dependendo da relação. Essa característica das relações CST, aliás, justifica o fato de que a maioria dos trabalhos que propõem métodos automáticos de identificação das relações CST são baseados na identificação da similaridade ou redundância entre as sentenças.

Quanto à complementaridade, a própria definição das relações CST que codificam esse fenômeno evidencia tal característica (cf. Seção 2.2). Assim, a similaridade ou redundância é uma das características que marcam as sentenças que compõem os pares anotados com as relações CST de complementaridade, como em (9).

(9) Sentença 1: A pesquisa foi realizada entre os dias 29 e 31 de julho e foi registrada no TSE com o número 12.197/2006.
Sentença 2: A pesquisa ouviu 2.002 pessoas entre os dias 29 e 31 de julho, em 142 municípios do país.

As sentenças em (9) veiculam informação sobre uma pesquisa de intenção de voto. Especificamente, ambas veiculam a informação sobre o período em que a pesquisa foi realizada, o que pode ser visto pela ocorrência dos trechos: "a pesquisa" e "entre os dias 29 e 31 de julho". Cada uma delas, no entanto, veiculam informações adicionais, que não estão contidas umas nas outras. No caso, a Sentença 1 veicula o número de registro da pesquisa no TSE e a Sentença 2, por sua vez, veicula a quantidade de eleitores ouvidos e de municípios que participaram da pesquisa.

b) A localização no texto-fonte

De acordo com Souza *et al*. (2012), dado um par de sentenças, a localização de cada umas delas em seus respectivos texto-fonte também é um relevante atributo que indica a similaridade entre elas. Segundo os autores, quanto menor for a distância entre as posições ocupadas pelas sentenças em seus respectivos textos-fonte, mais conteúdo em comum elas têm. Essa observação feita pelos autores apoia-se na estrutura típica dos textos jornalísticos. Segundo Lage (2002), um texto do tipo informativo é construído com base no método da pirâmide invertida (Figura 7).

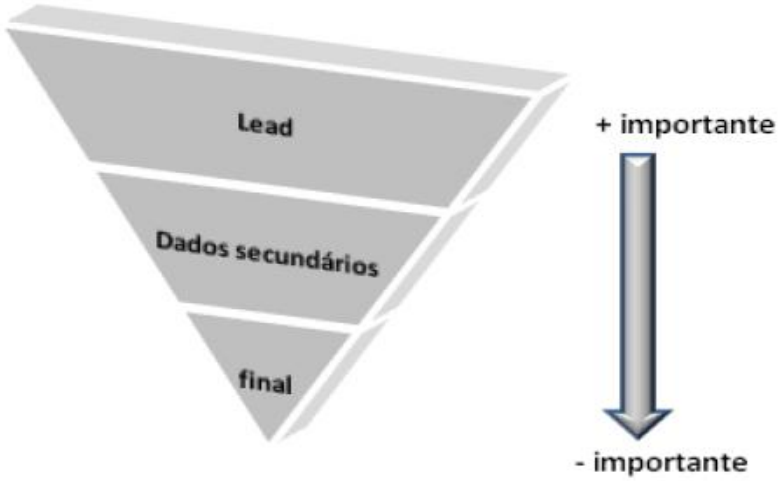

Figura 7: Estrutura do texto jornalístico segundo o método da pirâmide invertida.

Fonte: Lage (2002)

O método da pirâmide invertida ordena a informação de forma decrescente de relevância. O texto, então, é organizado em função de: (i) o *lead*, que corresponde ao primeiro ou aos dois primeiros parágrafos do texto e que expressa a informação principal a ser relatada, (ii) o corpo do texto, que desenvolve os elementos informativos referidos no *lead*, e (iii) o encerramento do texto (LAGE, 2002).

Dessa forma, caso as sentenças de um par tenham sido adquiridas do *lead* de seus respectivos textos-fonte, por exemplo, elas possivelmente compartilham informação ou conteúdo, pois são provenientes das mesmas regiões textuais.

De acordo com a anotação CST do *corpus* CSTNews, a Sentenças 1 do Quadro 15 e a Sentença 1 do Quadro 16 estão conectadas por uma relação CST de redundância (*Equivalence*), pois o conteúdo informativo de ambas é bastante semelhante. No caso, ambas veiculam: (i) quantas pessoas morreram no incidente no mercado de Moscou (“nove pessoas morreram, sendo três crianças”), (ii) quantos pessoas ficaram feridas (“25 ficaram feridas”), (iii) o dia (“segunda-feira”) e (iv) o local do incidente (“Moscou”).

Ao se calcular a similaridade entre elas com base na “sobreposição da localização”, verifica-se que a distância entre as posições ocupadas por ambas é 0 (zero), o que indica que elas são altamente similares quanto a esse atributo.

Quadro 15 – Texto 1

[1] 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. [2] A explosão, supostamente causada por vazamento de um botijão de gás, foi registrada por volta das 10h40 (3h40 de Brasília) no setor denominado Evrazia do mercado Cherkizov, um dos maiores shoppings da capital da Rússia. [3] A maioria dos feridos, entre os quais há quatro com menos de 18 anos, foi hospitalizada. [4] Cerca de dez de carros de bombeiros e mais de uma dezena de ambulâncias foram enviadas ao local, que foi isolado pela polícia.
[5] A procuradoria de Moscou anunciou a criação de um grupo especial para investigar o acidente. [6] Fontes do Ministério do Interior da Rússia citadas pela agência Interfax descartaram a possibilidade de a explosão em Cherkizov ter sido um ataque terrorista.

Fonte: http://143.107.232.31/CSTNews/source/D2_C15_Estadao.txt.seg/

Quadro 16 – Texto 2

[1] 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. [2] A explosão, cujas causas ainda são desconhecidas, aconteceu às 10h40 (3h40 em Brasília) no mercado Cherkizov, localizado no nordeste da capital russa. [3] A maioria dos feridos, entre os quais há quatro menores, foi hospitalizada. [4] A explosão - supostamente de um bujão de gás, segundo versões policiais preliminares - aconteceu no setor denominado "Evrazia" do mercado Cherkizov, um dos maiores shoppings da capital russa. [5] Cerca de dez carros de bombeiros e mais de uma dezena de ambulâncias foram enviadas ao local, que foi isolado pela Polícia.
[6] A procuradoria de Moscou anunciou a criação de um grupo especial para investigar o acidente. [7] Fontes do Ministério do Interior da Rússia citadas pela agência "Interfax" descartaram a possibilidade de a explosão em Cherkizov ter sido um ataque terrorista.

Fonte: http://143.107.232.31/CSTNews/source/D4_C15_JB.txt.seg/

Tendo em vista a relevância da localização no texto-fonte para identificar a similaridade ou redundância entre sentenças relacionadas, decidiu-se verificar como a complementaridade ocorre quanto a esse atributo. Para tanto, partiu-se da hipótese de que, na complementaridade, a distância entre as sentenças deve ser maior, pois, ao compartilharem menos conteúdo (que na redundância), as sentenças tendem a ocupar

posições mais distantes nos textos-fontes. Por exemplo, entre a Sentença 2 do Quadro 15 e a Sentença 7 do Quadro 16, que foram anotadas com a relação de complementaridade *Follow-up*, verifica-se que a distância entre a posição por elas ocupadas nos respectivos textos-fonte é 5 (cinco), pois as sentenças possuem conteúdo redundante (isto é, a explosão e o local do incidente), mas a Sentença 7 do Quadro 16 acrescenta a informação de que o governo russo descarta a possibilidade de a explosão ser um atentado terrorista.

c) O subtópico textual

Ainda quanto à estrutura do texto jornalístico, outra característica potencialmente relevante foi investigada. Esta, no caso, diz respeito ao tópico principal dos textos-fonte e aos subtópicos que os compõem.

Sabe-se que um texto jornalístico veicula um assunto (tópico) e detalhes sobre ele (subtópicos). De acordo com Koch e Elias (2009), esses detalhes constituem segmentos que ora se ligam diretamente, ora se ligam indiretamente ao assunto principal (tópico).

Assim, se a estrutura de "pirâmide invertida" dos textos jornalísticos diz respeito à organização informacional, logo também aborda a progressão topical. Assim, se duas sentenças que ocorrem em posições similares em seus textos-fonte são redundantes (SOUZA et al., 2012), é bastante provável que elas possuam identidade de subtópico.

Segundo a anotação de subtópicos do CSTNews realizada por Cardoso *et al.* (2012), o texto do Quadro 15 e do Quadro 16 estão segmentados em subtópicos, como indicado pelas linhas pontilhadas. Segundo essa segmentação, as Sentenças 1, 2, 3 e 4 do Quadro 15, por exemplo, veiculam um subtópico ("a explosão") e as Sentenças 5 e 6 veiculam outro subtópico ("o motivo da explosão"). Considerando-se a Sentença 1 do Quadro 15 e a Sentença 1 do Quadro 16, que foram anotadas com a relação CST *Equivalence*, observa-se que ambas possuem o mesmo subtópico ("a explosão").

Partindo-se do fato de que a complementaridade envolve certa redundância e também informação distinta entre as sentenças, verificou-se nos 135 pares iniciais do *subcorpus* como a complementaridade se relaciona com a expressão dos subtópicos.

Além dessas características gerais, ou seja, típicas de qualquer par de sentenças com complementaridade, a análise da parcela preliminar do *subcorpus* aqui apresentada permitiu que fossem identificados potenciais atributos específicos dos diferentes tipos de complementaridade (temporal e atemporal). A seguir, esses atributos são descritos.

3.1.2.2. Características específicas da complementaridade temporal

a) Os advérbios

Como mencionado, a informação complementar veiculada por uma sentença em relação e outra sentença pode envolver aspectos temporais. Sabendo-se que alguns advérbios têm a função de acrescentar ao verbo uma circunstância de tempo, os pares de sentenças do *subcorpus* anotados com relações CST de complementaridade temporal foram analisados com o objetivo de verificar se a informação complementar veiculada por uma das sentenças dos pares está ligada a advérbios de tempo, como exemplificado em (10).

(10) Sentença 1: A ofensiva israelense foi lançada depois de uma sequência de ataques do Hezbollah no domingo que causou as maiores baixas para Israel nas quatro semanas do conflito.

Sentença 2: Durante este domingo, dia 6, foram travadas lutas sangrentas.

Em (10), o advérbio "depois" está vinculado a uma informação adicional (isto é, "uma sequência de ataques do Hezbollah") e evidencia uma informação temporal sobre o evento principal ("a ofensiva israelense").

b) As expressões de tempo

Além dos advérbios, há em português algumas expressões que também indicam tempo. De acordo com Baptista *et al.* (2008), as expressões temporais (ETs) podem ser de 4 tipos: (i) tempo calendário, (ii) frequência (p.ex.: "Ocorrerá entre os dias 29 e 31 de julho"), (iii) duração (p.ex.: "O Natal é comemorado todo ano") e (iv) genérico (p.ex.: "Eu gosto do mês de julho"). As ETs que expressam "tempo calendário" podem ser de 3 subtipos: (i) hora (p.ex.: "Ele chegou às 9h30m), (ii) data e (iii) intervalo (p.ex.: "Entre junho e julho"). E, por fim, as ETs do subtipo "data" podem ser: (i) enunciação (p.ex.: "Partiu em março"), (ii) textual (p.ex.: "Um dia após a venda") e (iii) absoluto (p.ex.: "O acidente ocorreu em fevereiro de 2002").

Menezes Filho e Pardo (2011) utilizam a tipologia de Baptista *et al.* (2008) para identificar e anotar (ou seja, explicitar por meio de etiquetas identificadoras) ETs no *corpus* CSTNews. Menezes Filho e Pardo (2011) identificaram aproximadamente 1.000 ETs em todo o *corpus.* Em 0, têm-se exemplos de expressões em um par de sentenças complementares que foi anotado com a relação *Historical background.*

(11) Sentença 1: Acidentes aéreos são frequentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.

Sentença 2: 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.

Em (11), a Sentença 1 expressa uma informação histórica/passada sobre um elemento presente em Sentença 2. No caso, a Sentença 2 veicula a informação sobre um acidente aéreo no Congo e a Sentença 1 contém a informação de que esse tipo de acidente é comum nessa região.

Na Sentença 2, as expressões "na quinta-feira" e "nesta sexta-feira", que evidenciam quando o acidente ocorreu e quando a informação foi divulgada, foram anotadas como ETs. Segundo a tipologia em que se basearam, Menezes Filho e Pardo (2001) classificaram essas ETs como do tipo "data". Na Sentença 1, a expressão "à tarde" foi anotada como uma ET do subtipo "textual", já que indica textualmente (em comparação aos valores números) o período do dia em que ocorreu o evento.

3.1.2.3. Características específicas da complementaridade atemporal

A complementaridade atemporal parece não possuir muitas pistas ou marcas linguísticas na superfície textual como a complementaridade temporal. No entanto, alguns marcadores discursivos podem indicar a ocorrência de um detalhamento ou de uma informação adicional. Diz-se isso com base na utilização de tais marcadores para a identificação automática de relações semântico-discursivas monodocumentos. O analisador discursivo automático Dizer 2.0, por exemplo, construído por Mazeiro e Pardo (2011) para o português, é uma ferramenta de PLN que identifica as relações do modelo RST com base na ocorrência de alguns marcadores discursivos. No Quadro 17, apresenta-se o conjunto de marcadores utilizados pelo Dizer 2.0 para identificar, no caso, a relação *Elaboration* (atemporal) no cenário monodocumento, a qual é similar à relação CST de mesmo nome.

Quadro 17: Marcadores discursivos de complementaridade do Dizer 2.0.

Marcador	**Relação**	**Marcador**	**Relação**
adicionalmente	Atemporal	*desse modo*	Temporal
*ainda****	Atemporal	*em adição*	Atemporal
*além de****	Atemporal	*em comparação*	Atemporal
além disso	Atemporal	*em nível de*	Atemporal
analogamente	Atemporal	*em particular*	Atemporal
após	Temporal	*especificamente*	Atemporal
assim	Temporal	*essencialmente*	Atemporal
atualmente	Temporal	*inclusive*	Atemporal
bem como	Atemporal	*onde****	Atemporal
bem	Atemporal	*para retornar ao meu ponto*	Atemporal
com relação a	Atemporal	*por exemplo****	Atemporal
como exemplo	Atemporal	*por falar em*	Atemporal
como por exemplo	Atemporal	*realmente*	Atemporal
*(como) também****	Atemporal	*(sendo) assim****	Temporal
da mesma forma	Atemporal	*também*	Atemporal
de fato	Atemporal	*tanto que*	Atemporal
dessa forma	Temporal / Atemporal	*voltando ao assunto*	Atemporal

Fonte: Elaborado pelo autor.

Os atributos assinalados com (*) no Quadro 17 foram encontrados nos 135 pares de sentenças de complementaridade já caracterizados, em especial nos pares que possuem a relação *Elaboration*[24]. Em (12) exemplifica-se pares de sentenças em que alguns dos marcadores discursivos do Quadro 17 ocorrem.

(12)

A. Sentença 1: O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.

Sentença 2: Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.

[24] Apesar de Maziero e Pardo (2011) utilizarem os marcadores discursivos para identificar, dentre outras relações do modelo RST, inclusive *Elaboration*, na análise do *subcorpus* desta pesquisa observou-se que tais marcadores também podem ocorrer em pares de sentenças anotadas pela relação *Follow-up* e *Historical background*.

B. Sentença 1: Em nota enviada após a exibição da reportagem, a TAM afirma "que não teve registro de qualquer problema mecânico neste avião no dia 16 de julho".

Sentença 2: 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.

O par de sentenças em (12.A) informam sobre um acidente aéreo no Congo com vítimas fatais. A Sentença 2 acrescenta outra informação ao tópico principal (que o avião "levava uma carga de minerais"), introduzida pelo marcador "também". Já em (12.B) os pares de sentença também informam sobre um acidente aéreo, mas no Brasil. A Sentença 2 complementa com uma informação ("a aeronave da TAM continuou voando", com o reversor direito desligado") o tópico principal (não houve "registro de qualquer problema mecânico neste avião"), também introduzido por um amrcador discursivo ("ainda assim").

3.1.2.3. A complementaridade linguisticamente não marcada

Em (13), (14), (15) e (16), há pares de sentenças anotados com a relação *Elaboration*, nos quais a complementaridade se estabelece principalmente por meio do conhecimento de mundo, que permite realizar certas inferências e correlacionar as sentenças

(13)

Sentença 1: Ele não antecipou o volume de recursos nem onde serão aplicados.

Sentença 2: Lula disse que o critério para o investimento nas cidades será técnico, não partidário.

Em (13), observa-se que não há marcas linguísticas evidentes que se relacionam à temporalidade ou à atemporalidade. Ainda assim, é possível perceber a existência de complementaridade. No caso, a informação complementar na Sentença 2 consiste no critério técnico da aplicação dos recursos mencionados, na Sentença 1. Dessa forma, a informação complementar se dá por conhecimento de mundo (por exemplo, saber que "Lula" é "presidente, ou seja, pessoa com poder para destinar recursos", na Sentença 2)

e por inferências que o falante realiza (por exemplo, “ele”, na Sentença 1, equivale a “Lula’, na Sentença 2).

(14)

Sentença 1: Ele havia sido decretado pelo CGE devido ao risco de que novos alagamentos surgissem.

Sentença 2: A forte chuva em São Paulo complicava o trânsito na manhã desta segunda-feira, 16, e fez com que o Centro de Gerenciamento de Emergência (CGE) da Prefeitura colocasse a cidade em estado de atenção.

Em (14), não existem marcas linguísticas que explicitamente evidenciam a complementaridade entre as sentenças. Entretanto, a complementaridade é compreendida pelo falante por meio de correferências realizada com auxílio de conhecimento de mundo (“ele”, na Sentença 1, corresponde a “estado de atenção”, na Sentença 2). Dessa forma, a informação complementar, em (15), é o motivo do decreto do estado de alerta (no caso, “a forte chuva em São Paulo”), também realizada pelo falante por conhecimento de mundo.

(15)

Sentença 1: Algumas já estão em andamento, outras vão começar a andar agora, outras ainda precisam de licenciamento.

Sentença 2: "O nosso desejo, agora, é que essas obras que foram anunciadas agora, até fevereiro elas estejam licitadas e estejam gerando os empregos e a melhoria de vida que tanto nós precisamos para o nosso Brasil".

Em (15), a relação de complementaridade também se estabelece com base no mecanismo de correferência. Entretanto, a correferência é feita por relações indiretas e não linguisticamente marcadas. Vale ressaltar que na Sentença 1 o conhecimento de mundo que o falante emprega para compreensão desse conteúdo permite-o inferir que os pronomes indefinidos “algumas” e “outras” modificam algo que está elíptico (no caso, “as obras”). Assim, a informação complementar é “o desejo de que as obras sejam licitadas” gerando empregos e melhoria de vida”.

(16)

Sentença 1: A fase final da competição deste ano acontecerá na Rússia.

Sentença 2: O time está perto da classificação para a próxima fase.

Em (16), a Sentença 2 veicula uma informação adicional sobre a fase final de uma competição esportiva: a de que "o time está próximo de se classificar para jogá-la". Essa informação pode ser tida como complementar porque se entende, com base em conhecimento especializado (esportivo) do falante, o qual o permite inferir que "a próxima fase" de S2 é a "fase final da competição" veiculada por S1.

3.1.3. Caracterização do subcorpus

A caracterização dos 135 pares inicialmente analisados consistiu na descrição ou explicitação das características ou atributos linguísticos que tipificam a complementaridade (e a redundância) e que, por isso, têm o potencial de subsidiar a identificação automática do fenômeno em questão e de seus subtipos.

Para tanto, realizou-se a seleção e a delimitação dos atributos a partir das características levantadas anteriormente. Uma vez selecionados e delimitados, os atributos referentes aos 135 pares iniciais foram descritos e sua pertinência foi preliminarmente avaliada.

a) <u>Seleção e delimitação de atributos que tipificam a complementaridade</u>

No total, considerou-se um conjunto de 6 atributos, os quais estão descritos no Quadro 8. Os atributos selecionados podem ser divididos em: (i) atributos que detectam a similaridade ou redundância entre as sentenças que compõem um par e (ii) atributos típicos do fenômeno da complementaridade.

Quadro 18: Atributos para a caracterização da complementaridade.

Fenômeno	**Atributo**
Redundância	Sobreposição de nomes
	Distância
Complemenaridade	Sobreposição de advérbio
	Sobreposição de expressão temporal
	Sobreposição de subtópico
	Sobreposição de marcador discursivo

Fonte: Elaborado pelo autor.

Para capturar a redundância, foram selecionados dois atributos com base no trabalho de Souza *et al*. (2012). Dentre os vários atributos testados por Souza *et al*. (2012) para a detecção da redundância entre sentenças em português, a "distância" e a "sobreposição de nomes" são atributos simples de serem descritos e que capturam com alta precisão a similaridade entre sentenças de textos jornalísticos distintos sobre mesmo assunto.

O atributo "distância" também foi definido segundo as diretrizes de Souza *et al*. (2012). Assim, segundo esse atributo, quanto menor a distância entre as posições nos textos-fonte, maior a probabilidade de as sentenças serem similares. No exemplo em (18), as Sentenças 1 e 2, que estão conectadas pela relação CST *Historical background*, ocupam a mesma posição e, por isso, a diferença entre as posições das sentenças desse par é 0, o que revela que tendem a ter bastante conteúdo em comum.

O atributo "sobreposição de advérbio" busca codificar se as sentenças de um par possuem advérbios temporais envolvidos diretamente na complementaridade. Esse atributo foi subdivido em função da ocorrência do advérbio desse tipo de advérbio na Sentença 1 e na Sentença 2 do par. Assim, passou-a a ter dois atributos específicos, a saber: Adv_S1 e Adv_S2.

(17)

(A)

Sentença 1: Segundo fontes militares e policiais, os milicianos do Hisbolá já dispararam aproximadamente 2,7 mil foguetes Katyusha e mísseis de diferentes alcances contra território israelense desde de o início dos conflitos, que chega hoje ao seu 27º dia.

Sentença 2: Comandos israelenses mataram outros três guerrilheiros libaneses na cidade de Tiro, onde destruíram sete plataformas de lançamento de foguetes, informaram as fontes israelenses.

(B)

Sentença 1: Braço direito do peemedebista desde sua posse, ela é funcionária de carreira do BNDES e comandou a Secretaria de Previdência Complementar no governo Fernando Henrique Cardoso.

Sentença 2: Inicialmente, Solange Vieira, que é assessora especial de Jobim havia sido escolhida para comandar a Secretaria Nacional de Aviação Civil, a ser criada na estrutura do ministério, segundo a assessoria de imprensa do ministério.

O atributo "sobreposição de ET" busca capturar se as sentenças de um par possuem ETs envolvidas diretamente na complementaridade. Esse atributo foi subdivido em função da ocorrência das ETs na Sentença 1 e na Sentença 2 do par. Assim, passou-a a ter dois atributos específicos, a saber: ET_S1 e ET_S2.

(18)

A.

Sentença 1: Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas.

Sentença 2: A TAM confirmou, na noite desta quinta-feira, que ao *airbus* da TAM estava com o reverso do lado direito desligado, desde o último dia 13.

B.

Sentença 1: 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.

Sentença 2: A TAM confirmou, na noite desta quinta-feira, que ao *airbus* da TAM estava com o reverso do lado direito desligado, desde o último dia 13.

O atributo “sobreposição de subtópico” busca capturar se as sentenças de par possuem o mesmo subtópico textual ou não. Por exemplo, em (20), de acordo com a anotação de Cardoso et al. (2012), a Sentença 1 possui o subtópico “ataques do exército israelense”, enquanto que a Sentença 2 possui o subtópico “armamento bélico do Hezbolá”. Assim, as sentenças não mantêm sobreposição de subtópico. A hipótese desse atributo é a de que se o par de sentenças possui subtópicos distintos, é possível que as sentenças estejam sob alguma relação de complementaridade.

(19)

Sentença 1: Comandos israelenses mataram outros três guerrilheiros libaneses na cidade de Tiro, onde destruíram sete plataformas de lançamento de foguetes, informaram as fontes israelenses.

Sentença 2: Enquanto isso, soldados israelenses mataram 10 integrantes da milícia do Hezbollah.

O atributo “sobreposição de marcador discursivo” busca capturar se as sentenças de um par possuem marcadores discursivos envolvidos diretamente na complementaridade. Esse atributo foi subdivido em função da ocorrência dos marcadores discursivos na Sentença 1 e na Sentença 2 do par. Assim, passou-a a ter dois atributos específicos, a saber: MD_S1 e MD_S2. O uso desse atributo pauta-se na hipótese de que é possível identificar relações de complementaridade a partir do conjunto de marcadores utilizados por Maziero e Pardo (2011), ainda que para o modelo RST. Em (21) exemplifica-se um par de sentenças, em que a informação complementar na Sentença 2 é introduzida por um marcador discursivo (“ainda assim”).

(20)

Sentença 1: De acordo com a companhia aérea, a recomendação da Airbus -- fabricante do avião -- é que a revisão no reversor seja feita até dez dias depois de o defeito ser detectado.

Sentença 2: O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, e <u>ainda assim</u> a aeronave da TAM, um Airbus A320, continuou voando, com o reverso direito desligado.

Assim definidos, os atributos das sentenças dos 135 pares foram efetivamente descritos por meio de um processo semiautomático, demonstrado a seguir. Vale ressaltar que, quanto à descrição, não somente os 135 pares iniciais foram caracterizados, mas todos os pares que compõem o *subcorpus* extraído do CSTNews já passaram por esse processo.

b) Descrição semi(automática) dos atributos

Os atributos das sentenças dos 135 pares iniciais do *subcorpus* foram descritos (ou explicitados) de forma semiautomática para que, ao final, a pertinência dos mesmos quanto à tipificação do fenômeno pudesse ser avaliada.

Para o cálculo da "sobreposição de nome" e "sobreposição de advérbio", foi preciso descrever ou explicitar todas as palavras pertencentes a cada uma dessas classes de palavras que compunham as sentenças do *subcorpus.* Para tanto, o *subcorpus* passou por um pré-processamento, que consistiu na etiquetação morfossintática automática realizada por meio do *parser* PALAVRAS (BICK, 2000), considerado o de melhor desempenho para o português. Sendo um *paser*, ou seja, um analisador sintático automático, o PALAVRAS realiza a tarefa de etiquetação morfossintática, que é necessária à analise sintática. Assim, por meio dele, foi possível identificar automaticamente a classe das palavras que compõem o *subcorpus*, agilizando a caracterização ou explicitação das informações linguísticas necessárias à verificação dos atributos relativos aos nomes e advérbios.

Para o cálculo da "distância" entre as sentenças, a localização das mesmas em seus respectivos textos-fonte foi recuperada manualmente do *corpus* CSTNews, pois esta e outras informações sobre cada sentença do *corpus* estão disponíveis na interface *online* do *corpus*.

Para o cálculo da "sobreposição de ET", foi preciso explicitar as ETs de cada sentença do *subcorpus.* Tal informação foi recuperada manualmente da anotação prévia do CSTNews realizada por Menezes Filho e Pardo (2011) e também disponibilizada na interface *online* do *corpus*. Especificamente, foi recuperada, para cada sentença, a informação de que nela ocorria ou não ETs dos tipos ("frequência", "duração" e "genérico") e subtipos ("data", "hora" e "intervalo").

Para verificar a "sobreposição de subtópico", foi necessário explicitar o subtópico de cada sentença. Tal informação foi manualmente recuperada da anotação prévia do CSTNews realizada por Cardoso *et al*. (2012).

Para o cálculo da "sobreposição de marcador discursivo," explicitou-se manualmente a ocorrência ou não desses elementos na Sentença 1 e na Sentença 2 de cada par. Para tanto, utilizou-se a lista de marcadores discursivos utilizados por Maziero e Pardo (2011) na construção do analisador discursivo monodocumento Dizer 2.0 como guia para a descrição das sentenças.

No Quadro 19, tem-se exemplificado a descrição das informações linguísticas subjacentes a cada um dos atributos, a qual foi organizada e armazenada em uma tabela no formato *xlsx*. Salienta-se que, por disposição espacial, os atributos foram abreviados, a saber: Dist (distância), N (sobreposição de nome), Adv (sobreposição de advérbio), ETs (sobreposição de Expressões Temporais), SubT (subtópico), MD (sobreposição de marcadores discursivos). Nesse quadro, observa-se que cada linha equivale a uma sentença do *subcorpus.* Nas colunas, tem-se registrado, além das informações referentes ao par, *cluster* e relação CST de cada sentença, o conjunto das informações linguísticas subjacentes aos atributos, as quais foram descritas.

Especificamente, quanto a S1 do par 6 , observa-se que o seguinte conjunto de informações linguísticas foi descrito:

a) N(ome): porta-voz, avião, Soviet, Antonov-28, fabricação, propriedade, companhia, Trasept Congo, carga, mineral; os quais foram identificados por meio do PALAVRAS;
b) Adv(érbio): também; o qual foi identificado por meio do PALAVRAS, independentemente de indicar ou não temporalidade;
c) E(xpressão) T(emporal): nsa (isto é, "não se aplica"); indicando que a informação em questão não ocorre na sentença.
d) SubT(ópico): 1; esse valor indica que a S1 veicula o subtópico de número 1, ou seja, o primeiro expresso no texto-fonte;
e) M(arcador) D(icursivo): também; identificado manualmente a partir da lista de marcadores utilizada no Dizer 2.0.

Quadro 19: Exemplo da caracterização do subcorpus.

Corpus				**Atributos linguísticos**					
Par	**Cluster**	**Relação**	**Sentença**	**Dist**	**N**	**Adv**	**ET**	**SubT**	**MD**
6	1	*Follow-up*	S1	6	porta-voz, avião, Soviet, Antonov-28, fabricação, propriedade, companhia, Trasept Congo, carga, mineral	também	nsa	1	também
			S2	1	acidente, localidade, Bukavu, leste, República Democrática do Congo, RDC, pessoa, porta-voz, Nações Unidas	nsa	nsa	1	nsa
53	5	*Historical Background*	S1	2	Braço, peemedebista, posse, funcionária, carreira, BNDES, Secretaria de Previdência Complementar, governo, Fernando Henrique Cardoso	nsa	nsa	1	nsa
			S2	3	Solange, Vieira, assessor, Jobim, comandar, Secretaria Nacional, de Aviação Civil, estrutura, ministério, assessoria, imprensa, ministério	inicialmente	nsa	1	nsa
124	4	*Elaboration*	S1	1	CGE, Centro de Gerenciamento de Emergências, Prefeitura, São Paulo, ponto, alagamento, cidade	nsa	data, hora	1	nsa
			S2	1	chuva, São Paulo, trânsito, Centro de Gerenciamento de Emergência, CGE, Prefeitura, cidade, estado, atenção,	nsa	nsa	1	nsa

Fonte: Elaborado pelo autor.

A partir da explicitação das informações linguísticas ilustradas no Quadro 19, procedeu-se à verificação ou cálculo dos atributos entre as sentenças dos pares. Esse verificação, em especial, gerou os dados necessários para a análise preliminar manual da pertinência dos atributos delimitados quanto à caracterização do fenômeno da complementaridade.

c) Verificação dos atributos entre as sentenças dos pares

Nesta subseção, descreve-se como cada atributo foi efetivamente verificado ou calculado entre as sentenças de dado par.

Para o cálculo da "sobreposição de nome", optou-se pela medida *noun overlap* utilizada por Souza *et al*. (2012), por ser uma variação da medida clássica *word overlap*. Para ilustrar seu cálculo, considera-se o par 6 do Quadro 19. A Sentença 1 e a Sentença 2 desse par possuem apenas o nome "porta-voz" em comum, sendo que a Sentença 1 possui 10 nomes (porta-voz, avião, Soviet, Antonov-28, fabricação, propriedade,

companhia, Trasept Congo, carga, mineral) e a Sentença 2 possui 9 nomes (acidente, localidade, Bukavu, leste, República Democrática do Congo, RDC, pessoa, porta-voz, Nações Unidas). Aplicando a fórmula *noun overlap*, como descrita em (7) (pág. 43), obtém-se o resultado numérico aproximado de 0,05, indicando baixa redundância.

O cálculo do atributo "distância" também foi baseado em Souza et al. (2012). Os autores apontam que, dado um par de sentenças, a diferença entre as posições ocupadas pelas sentenças em seus respectivos textos-fonte, pode indicar similaridade, uma vez que, se próximas, podem vincular sobreposição de conteúdo. Considerando o par 6 do Quadro 19, a Sentença 1 e a Sentença 2 têm posições 6 e 1, respectivamente, em seus textos-fonte. Assim, a distância entre essa duas sentenças é 5.

Para o cálculo do atributo "sobreposição de advérbios" foi considerado se, dado um par de sentenças, havia algum advérbio que veiculasse algum aspecto temporal atrelado à complementaridade. Além disso, foram considerados advérbios temporais alocados na primeira e na segunda sentenças do par. Dessa forma, em relação ao par 5 do Quadro 19, a Sentença 2 possui um advérbio ("inicialmente") que evoca informação temporal.

O atributo "sobreposição de ETs" foi calculado com base na anotação realizada de Menezes Filho e Pardo (2011), disponível no *corpus* CSTNews. Nessa anotação, os autores disponibilizaram um conjunto de expressões que evidenciam tempo, com base na classificação proposta por Baptista et al. (2008). Além de evidenciar o aspecto temporal, as ETs deviam estar vinculadas à informação complementar do par de sentenças (se na Sentença 1, ou na Sentença 2). No par de sentenças 124 do Quadro 19, por exemplo, a Sentença 1 possui as ETs "dará" e "hora" vinculadas à informação complementar.

Para o cálculo do atributo "sobreposição de subtópicos" baseou-se na anotação de Cardoso et al. (2012). Observou-se se os pares de sentenças anotadas por relações de complementaridade possuiriam subtópios semelhantes. De acordo com a anotação de Cardoso et al. (2012), o par de sentenças 6 do Quadro 19 possui subtópicos idênticos.

O cálculo do atributo "marcadores discursivos" foi realizado se, dado um par de sentenças, algum dos marcadores discursivos utilizados por Maziero e Pardo (2011) para identificação de relações da teoria RST no cenário monodocumento, ocorriam na Sentença 1 ou na Sentença 2. O par de sentenças 6 do Quadro 19, por exemplo, possui o marcador discursivo "também" na Sentença 1.

Os resultados dessa verificação foram organizados em uma tabela também no formato *xlsx* para que pudessem ser analisados de forma manual e também automática. A Tabela 6 ilustra a organização e o armazenamento dos dados. Nessa tabela, ressalta-se que: (i) os atributos Dist e N têm valores numéricos, (ii) SubT é um atributo binário, cujos valores podem ser "sim" ou "não", os quais indicam, respectivamente, a presença ou a ausência da informação na sentença, e (iii) os demais atributos possuem valores indicados por sequências de caracteres que expressam a própria palavra que ocorreu na sentença (p.x.: *também*) ou o seu tipo (p.ex.: data_hora, não_temporal).

Tabela 6: Exemplo da caracterização do *subcorpus.*

Par	**Cluster**	**Relação**	**Dist**	**N**	**Adv_S1**	**Adv_S2**	**ET_S1**	**ET_S2**	**SubT**	**MD_S1**	**MD_S2**
6	1	Follow-up	5	0	n_temp	nsa	nsa	data	sim	também	nsa
53	5	Historical background	1	1	nsa	n_temp	nsa	nsa	sim	nsa	nsa
124	3	Elaboration	0	3	nsa	nsa	data_hora	data	sim	nsa	nsa

Fonte: Elaborado pelo autor.

3.1.4. Avaliação preliminar da pertinência dos atributos

A partir dos dados ilustrados na Tabela 6, resultantes da verificação ou cálculo dos atributos, procedeu-se a uma análise preliminar da pertinência dos atributos quanto à tipificação ou caracterização do fenômeno da complementaridade, em especial, quanto à caracterização das relações CST.

A análise manual da correlação entre os atributos e as relações CST teve início com o cálculo da média simples dos valores obtidos por cada atributo em função dos pares anotados por cada relação CST. Na sequência, verificou-se, na Tabela 6 completa, o número de pares que obtiveram valores iguais ou superiores à média simples.

Por exemplo, considerando-se o atributo "distância", a média simples identificada foi 6, ou seja, a diferença em média entre as posições ocupadas por cada uma das sentenças de um par em seu respectivo texto-fonte é de 6 sentenças. Assim, com base nos valores do atributo "distância" obtidos para os 135 pares inicialmente analisados, verificou-se que: (i) dos 45 pares anotados com *Follow-up*, 17 possuem o atributo "distância" igual ou maior que 6; (ii) dos 45 pares anotados com *Historical background*, 18 pares possuem o valor do atributo "distância" igual ou maior que 6, e (iii) dos 45 pares anotados com *Elaboration,* 18 também possuem o valor desse atributo igual ou superior à média 6. E,

assim, procedeu-se para cada um dos atributos. Ressalta-se que, no caso dos atributos de valores não-numéricos (p.ex.: sobreposição de advérbio), considera-se valores categóricos (“sim ou “não”, por exemplo).

Uma vez calculado o número de pares anotados com cada uma das relações CST que apresentaram valores iguais ou acima da média de cada atributo, buscou-se verificar manualmente os atributos que potencialmente melhor caracterizam a complementaridade e seus subtipos (ou relações CST).

Para analisar a pertinência dos atributos para a tipificação da complementaridade codificada pela relação *Follow-up*, por exemplo, somaram-se os valores dos atributos desta relação (17+23+5+1+18+18+21+10+2), dividindo-se pela quantidade de atributos (9), resultando na média simples (13); os atributos que obtiveram resultados iguais ou acima da média foram computados. Fez-se esse procedimento para todas as relações de complementaridade. Os atributos destacados em cinza na Tabela 7 são aqueles que, por meio dessa estratégia, obtiveram bons satisfatórios.

Tabela 7: Análise manual de atributos linguístico-estrutural em relação à complementaridade.

Atributo	**Relação CST**		
	Follow-up	**Historical background**	**Elaboration**
Distância	17/45	18/45	18/45
Sobreposição de nome	23/45	16/45	26/45
Advérbio em S1	5/45	13/45	7/45
Advérbio em S2	1/45	6/45	5/45
Expressão temporal em S1	18/45	35/45	8/45
Expressão temporal em S2	18/45	18/45	17/45
Sobreposição de subtópico	21/45	10/45	22/45
Sobreposição de marcador discursivo em S1	10/45	6/45	8/45
Sobreposição de marcador discursivo em S2	2/45	2/45	3/45

Fonte: Elaborado pelo autor.

Com base na Tabela 7, observar que:

a) os atributos “distância” e “sobreposição de nome” em S1” não parecem expressar a diferença de complementaridade, pois destaca-se para as 3 relações CST;

b) o atributo “sobreposição de ET em S1” parece evidenciar somente a diferença entre as relações de complementaridade temporal (*Follow-up* e *Historical background*) e atemporal (*Elaboration*);
c) o atributo “advérbios em S1” parece diferenciar a relação *Historical background* em relação às demais;
d) o atributo “sobreposição de subtópico” parece identificar somente as relações *Follow-up* e *Elaboration*;
e) os atributos “advérbio em S2”, “marcador discursivo em S1” e “marcador discursivo em S2” não foram pontuados, pois não obtiveram valores acima da média; isso pode demonstrar que tais atributos podem não identificar a complementaridade, a ponto de caracterizá-la.

Além da análise da pertinência dos atributos quanto à caracterização das relações CST de complementaridade, realizou-se a análise da pertinência dos mesmos quanto à tipificação dos tipos de complementaridade (temporal e atemporal).

Para analisar a pertinência dos atributos no que diz respeito à complementaridade temporal (*Follow-up*), por exemplo, somaram-se os valores dos atributos desta relação (17+23+5+1+18+18+21+10+2), dividindo-se pela quantidade de atributos (9), resultando na média simples (13); os atributos que obtiveram resultados iguais ou acima da média foram computados. Fez-se esse procedimento para todas as relações de complementaridade.

Para analisar a pertinência dos atributos no que diz respeito à complementaridade atemporal, na Tabela 8,, somaram-se os valores das porcentagens dos atributos desse tipo de complementaridade (38.8%+43.3%+20%+7.7%+58.80%+40%+34.4+17.70%+4.4%), dividindo-se pela quantidade de atributos (9), resultando na média simples (29.46%); os atributos que obtiveram resultados iguais ou acima da média foram computados. Fez-se esse procedimento para todas as relações de complementaridade. Os atributos destacados em cinza são aqueles que, por meio dessa estratégia, obtiveram bons satisfatórios.

Apresentam-se os resultados dessa análise na Tabela 8, de modo que o valor percentual indica a ocorrência do atributo em relação ao tipo de complementaridade.

Tabela 8: Análise manual de atributos linguístico-estrutural em relação aos tipos de complementaridade

Atributo	**Relação CST**	
	Temporal	**Atemporal**
Distância	38,8%	40%
Sobreposição de nome	43,3%	57,7%
Advérbio em S1	20%	15,5%
Advérbio em S2	7,7%	11,11%
Expressões temporais em S1	58,8%	17,7%
Expressões temporais em S2	40%	37,7%
Sobreposição de subtópicos	34,4%	48,8%
Marcadores discursivos em S1	17,7%	17,7%
Marcadores discursivos em S2	4,4%	6,60%

Fonte: Elaborado pelo autor.

Com base na Tabela 8, aponta-se:

a) os atributo “distância”, “sobreposição de nome” e “sobreposição de subtópicos”, *a priori*, parecem não expressar a diferença de complementaridade temporal e atemporal, pois todos os atributos apontados são bem avaliados em suas respectivas categorias de complementaridade;
b) apesar de os atributos destacados em (a) não representarem distinção dos tipos de complementaridade, os atributos “sobreposição de nome” e “sobreposição de subtópico” parecem ter uma discreta relevância para identificar a relação de complementaridade atemporal;
c) os atributos “advérbio em S1”, “advérbio em S2”, “marcador discursivo em S1” e “marcador discursivo em S2” parecem não apresentaram bom desempenho em identificar a complementaridade no *subcorpus*. O resultado demonstra que tais atributos podem não diferenciar os tipos de complementaridade.
d) o atributo “ET em S1”, *a priori*, parece ser o único a diferenciar os tipos de complementaridade.

A análise manual preliminar dos dados parece confirmar algumas das hipóteses deste trabalho:

- ***Hipótese 1:*** *atributos superficiais e profundos de detecção da redundância são pertinentes para a identificação da complementaridade, já que o conteúdo entre*

duas sentenças pode estar sob certa sobreposição em relação complementar. De acordo com a análise, todas as relações de complementaridade possuem, ainda que minimamente, informações em comum. Essa informação, por sua vez, concentra-se na classe dos nomes. Além disso, se a redundância for elevada em um par de sentenças de complementaridade, em geral, haverá identidade entre os subtópicos das sentenças.

- ***Hipótese 2:*** *a complementaridade pode se manifestar na superfície linguística, e essa manifestação pode ser capturada por atributos específicos que tem o potencial de subsidiar métodos automáticos de detecção desse fenômeno*. A análise de *corpus* permitiu indicar por quais atributos linguísticos a complementaridade se manifesta. Além disso, a análise também serviu de base para indicar manifestações do fenômeno multidocumento em que não é possível modelar automaticamente no cenário atual do PLN; ou seja, a complementaridade também pode se manifestar sem o uso de atributos linguísticos.

- ***Hipótese 3:*** *métodos de detecção da complementaridade podem capturam os diferentes tipos de complemento (temporais e atemporais).* De acordo com a análise, até o momento, os atributos "sobreposição de advérbio em S1", "sobreposição de advérbio em S2", "sobreposição de nome", "sobreposição de subtópico" e "distância" são capazes de diferenciar a complementaridade temporal, da atemporal, a saber: os atributos "distancia", "sobreposição de nomes", "sobreposição de ET em S1" e "sobreposição de ET em S2" e "sobreposição de subtópicos" se destacam na identificação da complementaridade temporal, enquanto que "distancia", "sobreposição de nomes", "sobreposição de ET em S2" e "sobreposição de subtópicos" destacam-se para a complementaridade atemporal.

- ***Hipótese 4:*** *métodos de detecção da complementaridade capturam as relações CST que expressam complemento (Historical background, Follow-up e Elaboration).*Segundo a análise manual, até o momento, os atributos "sobreposição de advérbio", "sobreposição de nome", "sobreposição de subtópico" e "distância" são capazes de diferenciar a complementaridade temporal, da atemporal, a saber: o atributo "sobreposição de ET" se destaca na identificação da complementaridade

temporal, enquanto que os atributos “distância”, “sobreposição de nomes” e “sobreposição de subtópico” destacam-se para a complementaridade atemporal.

Na próxima subseção, apresentam-se as etapas em andamento, a saber: identificação dos métodos de detecção da complementaridade, e etapas futuras.

3.2. Etapas em andamento

Nesta subseção apresentam-se as etapas em andamento desta pesquisa, em especial, a etapa de identificação manual de métodos de detecção da complementaridade análise manual a partir da caracterização linguístico-estrutural, descrita na subseção anterior.

3.2.1. Identificação de métodos de detecção da complementaridade

Essa tarefa consiste em identificar, com base nas sentenças caracterizadas, os atributos mais relevantes para a identificação da complementaridade. A identificação dos atributos mais relevantes poderá ser feita por análise automática. Quando manual, as sentenças serão analisadas visando eleger atributos linguísticos sobressalentes que possam indicar a complementaridade por meio de alguma regularidade. Quando automática, serão utilizados algoritmos de aprendizado de máquina para depreender padrões e/ou regras para identificar quais atributos (ou a combinação deles) que possuam bom desempenho na detecção da complementaridade.

3.3. Etapas futuras

Nesta subseção, apresentam-se as etapas futuras desta pesquisa, a saber, (i) estudo da correlação entre métodos e tipos de relação CST, (ii) estudo da correlação entre os métodos e as relações CST e (iii) a avaliação. Salienta-se que as tarefas descritas a seguir serão desenvolvidas de modo automático, em especial, utilizando algoritmos de aprendizado de máquina (AM) para a geração de regras capazes de discriminar os diferentes tipos e relações de complementaridade. O *subcorpus* será submetido a um ambiente de AM em que os dados resultantes da aplicação dos atributos sejam submetidos ao ambiente composto por algoritmos de aprendizagem advindos de diferentes abordagens da Inteligência Artificial. Os algoritmos implementados que serão utilizados serão capazes de analisar automaticamente os dados de entrada e aprender padrões estatisticamente relevantes, os quais são expressos por regras.

3.3.1. Estudo da correlação entre os métodos e os tipos de relação CST

A partir do conjunto de métodos levantados na tarefa anterior, pretende-se relacioná-los aos tipos de complementaridade (temporais e atemporais). Com base nesse estudo, pretende-se identificar os métodos que expressam mais adequadamente tais diferenças de complemento. Esse estudo pode ser realizado manualmente e automaticamente.

3.3.2. Estudo da correlação entre os métodos e as relações CST

De posse do conjunto de métodos levantados na tarefa de análise, pretende-se relacionar os atributos linguísticos às relações CST de complementaridade, a saber: *Historical background, Follow-up e Elabaration*. Com base nesse estudo, pretende-se identificar os métodos que expressam mais adequadamente essas relações. Salienta-se que essa tarefa pode ser realizada manualmente e automaticamente.

3.3.3. Avaliação

Esta tarefa consiste na aplicação dos métodos mais eficientes identificados na tarefas anteriores à parcela do *subcorpus* para teste. Essa tarefa delimita-se em três possibilidades, a saber: (i) utilizar o *subcorpus* para caracterização do fenômeno pode ser submetido ao ambiente de AM e obter-se regras para a avaliação da complementaridade; (ii) utilizar o subcorpus para a caracterização do fenômeno pode ser submetido ao ambiente de AM acrescido de uma outra parcela do corpus CSTNews; (iii) utilizar uma outra parcela ainda não caracterizada do *corpus* CSTNews. utilizadaEsta tarefa também engloba a geração de automática de sumários acrescidos de métodos de detecção da complementaridade.

4. CRONOGRAMA ATUALIZADO

No Quadro 20, apresenta-se o cronograma atualizado para a realização das etapas em andamento (Tarefas 5) e futuras (Tarefas 6, 7 e 8). O cronograma é dividido em meses e compreende o período de 6 meses até que a dissertação de mestrado seja defendida. Ressalta-se que as atividades estão dentro do prazo estabelecido.

Tarefa 5: Identificação de métodos de detecção da complementaridade
Tarefa 6: Estudo da correlação entre os métodos e os tipos de relação CST
Tarefa 7: Estudo da correção entre os métodos e as relações CST
Tarefa 8: Avaliação

Quadro 20 - Atualização do cronograma

	2014			2015	
Tarefas	**Out.**	**Nov.**	**Dez**	**Jan.**	**Fev.**
Tarefa 5					
Tarefa 6					
Tarefa 7					
Tarefa 8					
Redação da dissertação e defesa					

Fonte: Elaborado pelo autor.

5. CONSIDERAÇÕES FINAIS

Até o momento, realizou-se a caracterização linguística do fenômeno da complementaridade com base em *corpus* e a análise manual dos atributos levantados.

A caracterização permite apontar como o fenômeno se manifesta na superfície textual, com vistas à modelagem para a identificação automática.

De acordo com a análise manual, alguns atributos não obtiveram desempenho significativo quanto à identificação dos tipos e relações de complementaridade. Entretanto, ainda se pensa em submeter o conjunto de 135 casos ao ambiente de aprendizado de máquina e observar se, de fato, os atributos sobreposição de advérbios (em S1 e em S2) e sobreposição de marcadores discursivos (em S1 e em S2) não são pertinentes. Pensa-se dessa maneira porque enquanto os atributos, em uma análise manual, foram analisados individualmente, a análise automática permitirá levantar regras e/ou grupos de atributos pertinentes quanto à identificação automática da complementaridade.

As dificuldades da pesquisa se baseiam na ausência de trabalhos na área que identifiquem as relações CST além da redundância. Aqui, a caracterização da complementaridade também parte da similaridade que existe entre os pares de sentenças, e propõem-se outros atributos que possam ser pertinentes quanto a identificação da complementaridade.

Outra dificuldade em se modelar o fenômeno em questão em atributos é por uma característica da natureza do fenômeno: por vezes, como observado, a complementaridade se dá por inferências e/ou correferências realizadas pelo próprio falante. Dessa forma, a modelagem do fenômeno em atributos torna-se custosa, já que a própria área de PLN, em PB, avançou pouco quanto à identificação de correferências, e não se podem mapear as inferências realizadas pelo falante ao reconhecer informações complementares.

Nas etapas futuras, objetiva-se realizar uma avaliação automática dos atributos que gere um conjunto de heurísticas a partir das quais atributos pertinentes para a proposição de métodos automáticos de detecção da complementaridade em *corpus* sejam identificados. Além disso, a análise automática dos atributos linguísticos permitirá, de forma mais consistente, verificar a pertinência dos mesmos para subsidiar a identificação automática dos tipos de complementaridade (temporal ou atemporal) e

das relações CST que codificam esse fenômeno, ou seja, *Follow-up*, *Historical background* e *Elaboration*.

Objetiva-se também realizar avaliações acerca do desempenho dos atributos elencados na produção dos sumários.

Por fim, ressalta-se que esta pesquisa já proporcionou a publicação dos 3 trabalhos listados abaixo, os quais são frutos diretos e indiretos das atividades realizadas até o momento atual da pesquisa.

1. SOUZA, J.W.C.; DI-FELIPPO, A.; PARDO, T.A.S. **Em Busca de Métodos de Detecção da Complementaridade para a Sumarização Automática Multidocumento.** In: III Jornada de Descrição do Português - JDP, 2013, Fortaleza. Proceedings of IX Brazilian Symposium in Information and Human Language Technology and Collocated Events, 2013.

2. SOUZA, J.W.C; DI-FELIPPO, A. **O Corpus CSTNews e sua Complementaridade Temporal.** In: Toos and Resources for Automatically Processing Portuguese and Spanish – ToRPoEsp 2014, São Carlos. 2014.

3. RASSI, A.P.; ZACARIAS, A.C.I.; MAZIERO, E.G.; SOUZA, J.W.C.; DIAS, M.S.; CASTRO JORGE, M.L.R.; CARDOSO, P.C.F.; BALAGE FILHO, P.P.; CAMARGO, R.T.; AGOSTINI, V.; DI FELIPPO, A.; SENO, E.R.M.; RINO, L.H.M.; PARDO, T.A.S. **Anotação de Aspectos Textuais em Sumários do Córpus CSTNews.** Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 394. NILC-TR-13-01. São Carlos-SP, Outubro, 59p. 2013.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLAN, J. **Automatic hypertext link typing**. Proceedings of the the seventh ACM conference on Hypertext.. p. 42-52. 1996.

BAXENDALE, P. B. **Machine-made index for technical literature: an experiment.** IBM Journal of Research and Development, v. 2, n. 4, p. 354-361, 1958.

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 the Proceedings of the 3rd RST Brazilian Meeting, pp. 88-105. Cuiabá/MT, Brasil. 2011.

CARDOSO, P.C.F.; RASSI, A.P.; MAZIERO, E.G.; NÓBREGA, F.A.A.; SOUZA, J.W.C.; DIAS, M.S.; CASTRO JORGE, M.L.R.; BALAGE FILHO, P.P.; CAMARGO, R.T.; AGOSTINI, V.; DI FELIPPO, A.; RINO, L.H.M.; PARDO, T.A.S. **Anotação de Subtópicos do *Corpus* Multidocumento CSTNews.** Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 389. NILC-TR-12-07. São Carlos-SP, Junho, 18p. 2012.

EDMUNDSON, H. P. New methods in automatic extracting. **Journal of the ACM (JACM)**, v. 16, n. 2, p. 264-285, 1969.

ESKIN, E. **Towards multi-document summarization by reformulation: Progress and prospects.** In: National Conference on Artificial Inteligence. 1999, Florida. Proceedings... p. 453–460. 1999.

FELLBAUM, C (Ed.). **Wordnet: an electronic lexical database (Language, speech and communication).** Massachusetts: MIT Press, 1998.

GUPTA, V.; LEHAL, G. S. A survey of text summarization extractive techniques. **Journal of Emerging Technologies in Web Intelligence**, v. 2, n. 3, p. 258-268, 2010.

HATZIVASSILOGLOU, J. L.; KLAVANS J.L.; HOLCOMBE, M. **Simfinder: a flexible clustering tool for summarization.** In: NAACL AUTOMATIC

SUMMARIZATION WORKSHOP. Pittsburgh, PA, USA. Proceedings… Pittsburgh, 2001, p.9. 2001.

HIRSCHMAN, L.; MANI, I. **Evaluation**. In: Handbook of Computational Linguistics. MITKOV, R. (ed). Oxford University Press. 2003.

JOHNSON, F.C.; PAICE, C.D; BLACK, W.J.; NEAL, A.P. **The application of linguistic processing to automatic abstract generation.** 1993.

JURAFSKY, D.; JAMES, H. Speech and language processing an introduction to natural language processing, computational linguistics, and speech, 2000.

KOCH, I.V. ELIAS, V. M. **Ler e escrever: estratégias de produção textual.** São Paulo: Contexto, p. 102-125. 2009.

KUMAR, Y. J.; SALIM, N.; RAZA, B. **Cross-document structural relationship identification using supervised machine learning.** Applied Soft Computing, v. 12, n. 10, p. 3124-3131, 2012.

LAGE, N. **Estrutura da notícia**. Ática, 1987.

MACCARTNEY, B.; GRENAGER, T.; DE-MARNEFFE, M. C.; CER, D.; MANNING, C. D. **Learning to recognize features of valid textual entailments.** In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (pp. 41-48). Association for Computational Linguistics. 2006.

MANI, I.; BLOEDORN, E. Multi-document summarization by graph search and matching. **arXiv preprint cmp-lg/9712004**, 1997.

MANI, I.; MAYBURY, M. T. **Advances in automatic text summarization**. MIT Press, v. 293, 1999.

MANN, W. C.; THOMPSON, S. A. **Rhetorical structure theory: A theory of text organization.** University of Southern California, Information Sciences Institute, 1987.

MARSI, E.; KRAHMER, E. **Classification of semantic relations by humans and machines.** In: Proceedings of the ACL workshop on Empirical Modeling of Semantic Equivalence and Entailment. Association for Computational Linguistics, 2005. p. 1-6.

MARTINS, C.B.; PARDO, T.A.S.; ESPINA, A.P.; RINO, L.H.M. **Introdução à Sumarização Automática.** Série de Relatórios do NILC. NILC-TR-08-04. São Carlos-SP. In press. 2008.

MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. **Identifying multi-document relations.** In: International Workshop on Natural Language Processing and Cognitive Science. Funchal, Madeira. Proceedings... Funchal, 2010. p. 60-9. 2010.

MAZIERO, E.G.. **Identificação automática de relações multidocumento**. Tese de Doutorado. Universidade de São Paulo. 2012.

MAZIERO, E.G.; PARDO. T.A.S. **DiZer 2.0 – a Web Interface for Discourse Parsing.** In the Extended Activities Proceedings of the 9thInternational Conference on Computational Processing of Portuguese Language - PROPOR. April 27-30, Porto Alegre/RS, Brazil. 2010.

MCKEOWN, K. R. et al. **Tracking and summarizing news on a daily basis with Columbia's Newsblaster**. Proceedings of the second international conference on Human Language Technology Research. p. 280-285. 2002.

MENEZES FILHO, L.A. e PARDO, T.A.S. **Detecção de Expressões Temporais no Contexto de Sumarização Automática.** In the Proceedings of the 2nd STIL Student Workshop on Information and Human Language Technology, pp. 1-3. 24 a 25 de Outubro, Cuiabá/MT, Brasil. 2011.

NENKOVA, A. **Discourse factors in multi-document summarization**. Proceedings of the National Conference on Artificial Intelligence. p. 1654. 2005.

PAICE, C. D. **The automatic generation of literature abstracts:** an approach based on the identification of self-indicating phrases. Proceedings of the 3rd annual ACM conference on Research and development in information retrieval. p. 172-191. 1980.

PARDO, T.A.S; NUNES, M.G.V. **On the development and evaluation of a Brazilian Portuguese discourse parser**. Revista de Informática Teórica e Aplicada, v. 15, n. 2, p. 43-64, 2008.

RADEV, D. R. **A common theory of information fusion from multiple text sources step one: cross-document structure.** Proceedings of the 1st SIGdial workshop on Discourse and dialogue-Volume 10. p. 74-83. 2000.

RADEV, D. R.; MCKEOWN, K. R. **Generating natural language summaries from multiple on-line sources.** Computational Linguistics, v. 24, n. 3, p. 470-500, 1998.

SAGGION, H.; LAPALME, G. **Concept identification and presentation in the context of technical text summarization.** Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization. p. 1-10. 2000.

SOUZA, J. W. C.; DI-FELIPPO, A.; PARDO, T. A. S. Investigação de métodos de identificação de redundância para Sumarização Automática Multidocumento. Série de Relatórios do NILC. NILC-TR-12. São Carlos-SP. In press. 2012.

SPARCK-JONES, K. ; GALLIERS, J. R. **Evaluating natural language processing systems:** An analysis and review. Springer, v. 1083, 1996.

SPARCK-JONES, K. **Automatic summarising: a review and discussion of the state of the art.** University of Cambridge Computer Laboratory, 2007.

TRIGG, R. **A Network-Based Approach to Text Handling for the Online Scientific Community**. PhD. Thesis. University of Maryland Technical Report, TR- 1346. College Park MD. 1983.

TRIGG, R. H.; WEISER, M. **TEXTNET: a network-based approach to text handling.** ACM Transactions on Information Systems (TOIS), v. 4, n. 1, p. 1-23. 1986.

VAN HALTEREN, H.; TEUFEL, S. **Examining the consensus between human summaries: initial experiments with factoid analysis.** Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5. p. 57-64. 2003.

VOSSEN, P. **Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-LingualIndex.** International Journal of Lexicography, v. 17, n. 2, p. 161-173, 2004.

WHITE, J. S.; DOYON, J. B.; TALBOTT, S. W. **Task tolerance of MT output in Integrated Text Processes.** ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems. p. 9-16. 2000.

ZHANG, Z.; GOLDENSHON, S.B.; RADEV, D.R. **Towards CST-Enhanced Sumarization.** In Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-2002). Edmonton/Canadá. 2002.

ZHANG, Z.; RADEV, D. **Combining labeled and unlabeled data for learning cross-document structural relationships.** In: Natural Language Processing – I JCNLP 2004. Springer. p. 32-41. 2005.

ANEXO 1 – Exemplo da relação de atributos

PAR	CLUSTER	RELAÇÃO	SENTENÇA	DISTÂNCIA	NOMES	ADVÉRBIOS	EXPRESSÕES TEMPORAIS	SUBTÓPICO	MARCADORES
1	1	Follow-up	S1	4	avião, porta-voz, ONU, Kinshasa Jean-Tobias, Okala	nsa	nsa	1	nsa
			S2	2	porta-voz, ONU, avião, fabricação, aeroporto, Bukavu, tempestade	nsa	nsa	1	nsa
2	1	Follow-up	S1	6	porta-voz, avião, Soviet, Antonov-28, fabricação, propriedade, companhia, Trasept Congo, carga, mineral	também	nsa	3	também
			S2	1	acidente, localidade, Bukavu, leste, República Democrática do Congo, RDC, pessoa, porta-voz, Nações Unidas	nsa	nsa	1	nsa
3	1	Follow-up	S1	5	sobrevivente, Okala	não	nsa	1	nsa
			S2	5	avião, porta-voz, ONU, Kinshasa Jean-Tobias, Okala	nsa	nsa	1	nsa
4	1	Follow-up	S1	6	porta-voz, avião, Soviet, Antonov-28, fabricação, propriedade, companhia, Trasept Congo, carga, mineral	também	nsa	3	também
			S2	4	avião, porta-voz, ONU, Kinshasa Jean-Tobias, Okala	nsa	nsa	1	nsa
5	1	Follow-up	S1	5	sobrevivente, Okala	não	nsa	1	nsa
			S2	4	avião, porta-voz, ONU, Kinshasa Jean-Tobias, Okala	nsa	nsa	1	nsa
48	3	Historical Background	S1	13	falha, reverso, causa, acidente, Fokker-100, TAM, decolagem, Congonhas	nsa	data	5	também
			S2	1	TAM, noite, airbus, reverso, lado, direito	nsa	nsa	1	nsa

49	3	Historical Background	S1	13	falha, reverso, causa, acidente, Fokker-100, TAM, decolagem, Congonhas	nsa	data	3	também
			S2	5	problema, detectado, sistema eletrônico, checagem, avião, aeronave, TAM, Airbus A320, reverso, direito	nsa	nsa	1	nsa
50	4	Historical Background	S1	17	julho, ano, passado, média, horário	nsa	data	1	nsa
			S2	2	CET, Companhia de Engenharia de Tráfego, congestionamento, cidade, média, horário,	nsa	nsa	1	nsa
51	4	Historical Background	S1	17	julho, ano, passado, média, horário	nsa	data	1	nsa
			S2	3	congestionamento, extensão, média	ainda_maior	nsa	1	nsa
52	4	Historical Background	S1	17	julho, ano, passado, média, horário	nsa	data	1	nsa
			S2	2	cidade, lentidão, média, horário, Companhia de Engenharia de Tráfego, CET	nsa	nsa	1	nsa
91	1	Elaboration	S1	4	avião, porta-voz, ONU, Kinshasa Jean-Tobias, Okala	nsa	nsa	1	nsa
			S2	1	pessoa, queda, avião, passageiro, República Democrática do Congo	nsa	nsa	1	nsa
92	1	Elaboration	S1	6	porta-voz, avião, Soviet Antonov-28, fabricação, propriedade, companhia, congolesa, Trasept, Congo, carga	nsa	nsa	3	nsa
			S2	1	pessoa, queda, avião, passageiro, República Democrática do Congo	nsa	nsa	1	nsa
93	1	Elaboration	S1	2	porta-voz, ONU, avião, fabricação,	nsa	data	1	nsa

					russa, aeroporto, Bukavu, tempestade				
			S2	1	acidente, localidade, Bukavu, leste, República Democrática do Congo, pessoa, porta-voz, Nações Unidas	hoje	nsa	1	nsa
94	1	Elaboration	S1	3	avião, tempo, pista, aterrissagem, floresta, aeroporto, Bukavu	não	nsa	1	nsa
			S2	2	porta-voz, ONU, avião, fabricação, russa, aeroporto, Bukavu, tempestade	nsa	nsa	1	nsa
95	1	Elaboration	S1	5	avião, acidentado, operado, Air Traset, passageiro, tripulante	nsa	nsa	1	nsa
			S2	4	avião, porta-voz, ONU, Kinshasa Jean-Tobias, Okala	nsa	nsa	1	nsa

ANEXO 2 – Exemplos de pares de sentenças com relação de complementaridade

PAR	RELAÇÃO	SENTENÇA
1	Follow-up	O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.
		Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.
2	Follow-up	O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.
		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.
3	Follow-up	"Não houve sobreviventes", disse Okala.
		O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.
4	Follow-up	O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.
		O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.
5	Follow-up	"Não houve sobreviventes", disse Okala.
		O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.
		Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 quilômetros.
48	Historical background	Acidentes aéreos são freqüentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.
		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.
49	Historical background	Aviões são usados extensivamente para transporte na República Democrática do Congo, um vasto país no qual há poucas estradas pavimentadas.
		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.
50	Historical background	Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas.
		A TAM confirmou, na noite desta quinta-feira, que ao airbus da TAM estava com o reverso do lado direito desligado, desde o último dia 13.
51	Historical background	Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas.
		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.
52	Historical background	Em julho do ano passado, a média foi de 36 km no horário.
		Naquele horário, segundo a CET (Companhia de Engenharia de Tráfego), havia 110 km de congestionamento em toda a cidade enquanto a média para o horário era de 76 km.
91	Elaboration	O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-

		Tobias Okala.
		Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.
92	Elaboration	O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.
		Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.
93	Elaboration	Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.
		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.
94	Elaboration	Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu.
		Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.
95	Elaboration	O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes.
		O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.