Programa de Pós-Graduação em Linguística

PPGL

DESCRIÇÃO LINGUÍSTICA DA COMPLEMENTARIDADE PARA A SUMARIZAÇÃO AUTOMÁTICA MULTIDOCUMENTO

JACKSON WILKE DA CRUZ SOUZA

SÃO CARLOS
2015

Universidade Federal de São Carlos

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS

PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

DESCRIÇÃO LINGUÍSTICA DA COMPLEMENTARIDADE PARA A SUMARIZAÇÃO AUTOMÁTICA MULTIDOCUMENTO

JACKSON WILKE DA CRUZ SOUZA

BOLSISTA: FAPESP (2013/21135-1)

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Linguística, área de concentração: Descrição, análise e processamento automático de línguas naturais.

Orientadora: Profa. Dra. Ariani Di Felippo

São Carlos - SP
2015

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária UFSCar
Processamento Técnico
com os dados fornecidos pelo(a) autor(a)

Souza, Jackson Wilke da Cruz
S729d Descrição linguística da complementaridade para a sumarização automática multidocumento / Jackson Wilke da Cruz Souza. -- São Carlos : UFSCar, 2016.
105 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2015.

1. Complementaridade. 2. Relações CST. 3. Linguística textual. 4. Descrição linguística. 5. Sumarização automática multidocumento. I. Título.

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS

PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

DESCRIÇÃO LINGUÍSTICA DA COMPLEMENTARIDADE PARA A SUMARIZAÇÃO AUTOMÁTICA MULTIDOCUMENTO

JACKSON WILKE DA CRUZ SOUZA

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Linguística, área de concentração: Descrição, análise e processamento automático de línguas naturais.

Membros da Banca:

(Orientadora – DL – UFSCar)

(Banca examinadora – DL – UFSCar)

(Banca examinadora – ICMC – USP/São Carlos)

São Carlos - SP
Março, 2015

Agradecimentos

Agradeço ao **Pai celeste** que me acompanhou todos os dias dessa jornada repleta de desafios, lágrimas, vitórias e alegrias: a Ti, Senhor, minha eterna gratidão que não pode ser expressa em palavras, mas em uma vida inteira a Ti – o que ainda seria pouco...

Agradeço à minha família (**José, Marlene e Jéssica Souza**): se cheguei até aqui, foi pelo esforço de nos mantermos juntos, ainda que as ondas do mar da vida nos mostrasse o contrário. Obrigado por serem os maiores incentivadores para que eu continuasse a percorrer o caminho acadêmico.

Muito obrigado à professora **Ariani Di Felippo** por, novamente, acreditar em mim e nessa parceria que já atravessa os anos. Ainda me lembro da primeira vez que sentamos para conversar sobre PLN: a paixão e o desejo por novas descobertas nesse universo linguístico-computacional se mantêm tão vivos quanto naquele dia. Muito obrigado!

Agradeço à **Profª. Drª. Gladis Maria de Barcellos Almeida** e ao **Prof. Dr. Thiago Alexandre Salgueiro Pardo** pela disponibilidade em participar junto a este trabalho desde a Qualificação (e sempre): sem dúvida alguma, o olhar de ambos enriqueceu a pesquisa.

Aos meus amigos que me acompanharam desde o começo dessa viagem, em especial **Luciana Rugoni** e **Marco Antônio Ruiz**: a cada encontro com vocês eu revivia momentos ainda de graduando, e vivia a esperança de uma vida acadêmica promissora. Muito obrigado pelos conselhos e motivações.

Aos meus **amigos que a vivência em São Carlos me presenteou**: muito obrigado por serem a presença de uma família aqui (vocês são muitos!). Obrigado por me permitirem entrar na vida de cada um de vocês e ser marcado pela presença de vocês no meu dia a dia.

Às minhas companheiras de lutas e vitórias **Roana Rodrigues** e **Ana Paula Cavaguti**: obrigado por sempre me incentivarem, por sempre trazerem palavras de calma e razão; por se fazerem de degraus para que eu crescesse. Muito obrigado.

Aos meus amigos paulistas e mineiros, em especial **Kaio**, **Flávio**, **Luís**, **Bruna**, **Fernanda** e **Michelle**: muito obrigado pela paciência, amizade e parceria de vocês. Encontrá-los sempre é ter a certeza que muitas coisas mudam, mas amizades verdadeiras perduram por uma eternidade.

À minha amada **igreja Missão Atos** onde fiz amigos que levarei pela vida toda: muito obrigado por me acolherem e me amarem em momentos ruins e por acompanharem meu

crescimento. Sou eternamente grato a Deus pela vida de cada um de vocês, em especial aos pastores **Claudio e Ana Ribeiro**: como é especial ser pastoreado por vocês.

Aos meus amigos e parceiros do NILC, em especial **Paula**, **Fernando**, **Lucía**, **Lianet**, **Márcio, André** e **Alessandro**: muito obrigado por sempre estarem dispostos a me esclarecerem dúvidas (sobretudo, computacionais) e a me ajudarem em pequenas e grandes tarefas!

Aos **professores do Departamento de Letras** e **do grupo de estudo SUCINTO**: continuo com a certeza de que sou fruto do trabalho e esforço conjunto de todos vocês.

Aos **funcionários da UFSCar** e **USP – São Carlos**: sem vocês muitos trabalhos seriam impedidos de serem realizados.

Às agências **Capes** e **FAPESP**: obrigado pelo suporte financeiro.

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 fontes distintas. Para tanto, a SAM seleciona informações mais importantes da coleção para compor o sumário. A seleção do conteúdo principal requer, por vezes, a identificação da redundância, complementaridade e contradição, que se caracterizam por serem os fenômenos multidocumento. A identificação da complementaridade, em especial, é relevante porque uma informação pode ser selecionada para o sumário uma vez que complementa outra já selecionada, garantindo mais coerência e informatividade. Alguns métodos de SAM realizam a condensação do conteúdo dos textos-fonte com base na identificação das relações do modelo/teoria *Cross Document Structure Theory* (CST) que se estabelecem entre as sentenças dos diferentes textos-fonte. Algumas dessas relações (p.ex., *Historical background*) capturam o fenômeno da complementaridade. A detecção automática dessas relações é comumente feita com base na similaridade lexical entre as sentenças, posto que as pesquisas sobre SAM não contam com estudos que tenham caracterizado o fenômeno, evidenciado outras estratégias linguísticas relevantes para detectar automaticamente a complementaridade. Neste trabalho, fez-se a descrição linguística da complementaridade com base em *corpus*, traduzindo as características desse fenômeno em atributos que subsidiam a sua identificação automática. Como resultados, obtiveram-se conjuntos de regras que evidenciam os atributos mais relevantes para a discriminação das relações CST de complementaridade (*Historical background*, *Follow-up* e *Elaboration*) e dos tipos (temporal e atemporal) da complementaridade. Com isso, espera-se contribuir para a Linguística Descritiva, com o levantamento baseados em *corpus* das características linguísticas do referido fenômeno, quanto para o Processamento Automático de Línguas Naturais, por meio das regras que podem subsidiar a identificação automática das relações CST e dos tipos de complementaridade.

Palavras-chave: complementaridade, relações CST, linguística textual, descrição linguística, sumarização automática multidocumento.

ABSTRACT

Automatic Multidocument Summarizarion (AMS) is a computational alternative to process the large quantity of information available online. In AMS, we try to automatically generate a single coherent and cohesive summary from a set of documents which have same subject, each these documents are originate from different sources. Furthermore, some methods of AMS select the most important information from the collection to compose the summary. The selection of main content sometimes requires the identification of redundancy, complementarity and contradiction, characterized by being the multidocument phenomena. The identification of complementarity, in particular, is relevant inasmuch as some information may be selected to the summary as a complement of another information that was already selected, ensuring more coherence and most informative. Some AMS methods to condense the content of the documents based on the identification of relations from the Cross-document Structure Theory (CST), which is established between sentences of different documents. These relationships (for example Historical background) capture the phenomenon of complementarity. Automatic detection of these relationships is often made based on lexical similarity between a pair of sentences, since research on AMS not count on studies that have characterized the phenomenon and show other relevant linguistic strategies to automatically detect the complementarity. In this work, we present the linguistic description of complementarity based on corpus. In addition, we elaborate the characteristics of this phenomenon in attributes that support the automatic identification. As a result, we obtained sets of rules that demonstrate the most relevant attributes for complementary CST relations (Historical background, Follow-up and Elaboration) and its types (temporal and timeless) complementarity. According this, we hope to contribute to the Descriptive Linguistics, with survey-based corpus of linguistic characteristics of this phenomenon, as of Automatic Processing of Natural Languages, by means of rules that can support the automatic identification of CST relations and types complementarity.

Keywords: complementarity, CST relations, textual linguistics, linguistic description, automatic multidocument summarizarion

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

Sumário

Capítulo 1

INTRODUÇÃO

1.1 Contextualização

O acesso e disponibilização da informação, nos dias atuais, estão cada vez mais fáceis e abundantes. De acordo com as projeções de Taufer (2013) para o ano de 2015, chegaria a ser produzido 8 *zetabytes* de informação e disponibilizado na Web. Para 2020, prevê-se que essa quantidade será de 40 *zetabytes* (TAUFER, 2013). Isso deixa o usuário frente a uma fonte quase que inesgotável de conhecimento.

Muitas subáreas do Processamento Automático de Línguas Naturais (PLN) estudam meios de lidar com essa vasta quantidade de informação, tais como a recuperação de informação, sistemas de pergunta e respostas e tradução automática. Além dessas subáreas, destaca-se a Sumarização Automática Multidocumento (SAM), na qual se objetiva automatizar a produção de sumários a partir de uma coleção de textos-fonte, advindos de fontes distintas, que abordam um mesmo assunto (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) (KUMAR, SALIM, 2012). Os 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: Adaptado de 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).

Mani (2001) aponta que 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.

Ainda de acordo com Mani (2001) 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, evitando-se que o sumário contenha informações redundantes e contraditórias, além de identificar as informações que são complementares umas às outras dentro da coleção. Têm-se, assim, os chamados fenômenos multidocumentos (redundância, contradição e complementaridade), pois resultam da multiplicidade de textos-fonte.

Dessa forma, a necessidade de identificação (e tratamento) dos fenômenos multidocumento ocorre porque: (i) as sentenças mais redundantes na coleção veiculam suas principais informações e, por isso, devem constar do sumário; (ii) as sentenças com conteúdo

complementar já selecionadas podem compor o sumário, e (iii) as sentenças redundantes ou contraditórias entre si não podem ser selecionadas para o sumário.

Por conseguinte, é preciso identificar, na fase de análise, os fenômenos de conteúdo típicos da multiplicidade de textos-fonte, sobretudo, os jornalísticos. Tais fenômenos, ou seja, a redundância, a complementaridade e a contradição, 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 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 expressam conteúdos que se contradizem (a Sentença 2 aponta que os candidatos em questão não pontuaram, e na Sentença 1 os mesmos candidatos obtiveram pontuações muito baixa).

A literatura propõe vários modelos para analisar segmentos textuais (TRIGG, 1983; TRIGG, WEISER, 1986; ALLAN, 1996; RADEV, MACKEOWN, 1998; AFANTENOS *et al.*, 2004; DAGAN *et al.*, 2009) em que se manifeste algum dos fenômenos muldicumento apresentados. De acordo com Taboada e Das (2013), modelos teóricos que se propõem a

analisar os relacionamentos existentes entre segmentos textuais centram-se no nível discursivo, observando a maneira como se organizam as unidades que compõem o discurso (logo, suas relações de coerência). Dessa forma, as autoras apontam que tais teorias têm o mesmo propósito: identificar relações possíveis entre proposições que constroem blocos discursivos, além de serem capazes de explicar a coerência existente dentro do discurso.

A Figura 2 ilustra um esquema genérico de relacionamento entre sentenças de textos que abordam mesmo tópico (assunto principal).

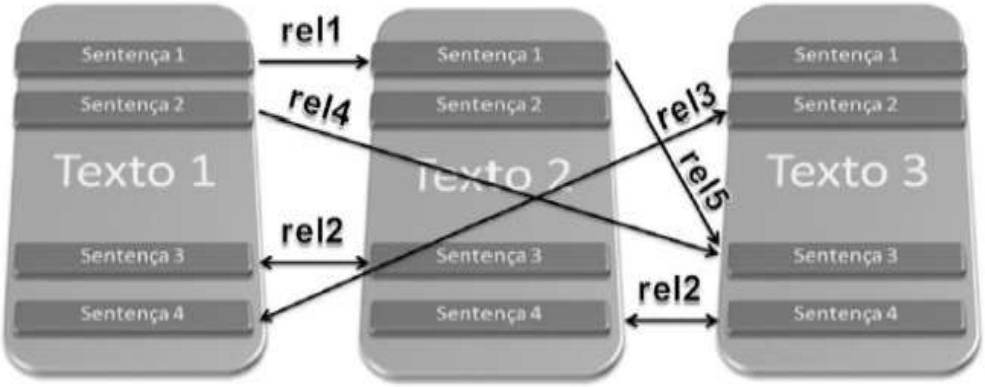

Figura 2: Esquema genérico de análise multidocumento.

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 de relações CST de Maziero *et al.* (2010) para o Português do Brasil (PB). Os autores agruparam esses fenômenos multidocumento em função da anotação manual do *corpus* CSTNews (CARDOSO *et al.*, 2011).

Quadro 1: Conjunto 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 do Quadro 1 capturam a "complementaridade" entre sentenças de um par. Entende-se complementaridade como a relação que se estabelece entre duas sentenças, S1 e S2, 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 via CST, o que equivale a 43% das relações.

Ainda de acordo com a tipologia de Maziero *et al.* (2010), a complementaridade pode ser temporal ou atemporal. A complementaridade temporal pode ser de dois tipos diferentes. Dado um par de sentenças S1 e S2, as sentenças 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 devem ter um espaço de tempo relativamente curto entre si (no modelo CST, essa relação é rotulada como *Follow-up*). Em (4) e (5), têm-se exemplos das relações de complementaridade temporal definida em (i) e (ii).

(4)

Sentença 1: O acidente ocorreu no delta do Nilo, ao norte de Cairo, no Egito.

Sentença 2: A maior tragédia ferroviária da história do Egito ocorreu em fevereiro de 2002, após o incêndio de um trem que cobria o trajeto entre Cairo e Luxor (sul), lotado de passageiros, e que deixou 376 mortos, segundo números oficiais.

(5)

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 (4), o par de sentenças veicula informação sobre um acidente ferroviário que ocorreu no Cairo, capital do Egito. A relação que há entre as sentenças do par é de *Historical background*, já que a Sentença 2 apresenta um fato histórico ("A maior tragédia ferroviária da

história do Egito ocorreu em fevereiro de 2002") relativo ao tópico principal veiculado pela Sentença 1 ("O acidente ocorreu no delta do Nilo, ao norte de Cairo, no Egito"). Em (5), o par de sentenças foi anotado com a relação *Follow-up*, e informa sobre um ataque israelense à milícia do Hezbollah. Na Sentença 2, aponta-se que as "lutas sangrentas" foram travadas no domingo, fato que sucedeu à ofensiva israelense ao Hezbollah após "quatro semanas de conflito".

A complementaridade atemporal 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.* De acordo com a definição, essa relação não envolve conteúdo que indica a localização no tempo (anterior ou posterior) de um acontecimento/fato com relação a outro. As sentenças em (6) ilustram esse tipo de complementaridade.

(6)

Sentença 1: 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.

Sentença 2: O Corpo de Bombeiros de São Paulo registrava apenas uma ocorrência grave ligada à chuva, às 9h30.

Em (6), o par de sentenças veicula informação sobre uma forte chuva em São Paulo e suas consequências no trânsito. A Sentença 2, com relação ao tópico principal, acrescenta que somente uma "ocorrência grave ligada à chuva foi registrada", a qual elabora o tópico principal.

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 (o tamanho desejado do sumário).

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.*, 2002, ZHANG, RADEV, 2005, MAZIERO et al, 2010).

Para o PB, o CSTParser (MAZIERO, PARDO, 2012) identifica as relações CST com precisão aproximada de 70%, baseando-se em atributos similares aos de Zhang *et al.* (2002) 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, HENDRICKX *et al.*, 2009, KUMAR *et al.*, 2012, SOUZA *et al.*, 2012).

Do que foi exposto, observa-se que: (i) a complementaridade é identificada 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, já que esse fenômeno não foi sistematicamente investigado.

Dessa forma, buscando melhorar a identificação automática da complementaridade, foram propostos os objetivos descritos na próxima subseção.

1.2 Objetivos e hipóteses

O objetivo desta pesquisa foi investigar o fenômeno da complementaridade circunstanciado pelo modelo CST visando suas aplicações na SAM. Objetivou-se realizar uma descrição linguística desse fenômeno multidocumento e propor, então, métodos de identificação automática do fenômeno em PB.

Assim, os seguintes objetivos específicos foram traçados:

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 automaticamente os diferentes tipos de complementaridade (temporal e atemporal) e as relações CST que os codificam (*Historical background*, *Follow-up* e *Elaboration*).

c) analisar a pertinência dos atributos por meio do paradigma supervisionado de Aprendizado de Máquina (AM), cujos algoritmos adquirem conhecimento implícito de exemplos previamente classificados, gerando classificadores (p.ex.: conjunto de regras) que relacionam os atributos (e seus valores) às classes.

Tais objetivos, aliás, relacionam-se diretamente à meta de pesquisa de um projeto maior denominado SUSTENTO[1] (FAPESP 2012/13246-5/ CNPq 483231/2012-6)[2], que é a de produzir e/ou sistematizar conhecimento linguístico para subsidiar a SAM do PB.

Os objetivos deste trabalho pautaram-se em 4 hipóteses sobre o fenômeno da complementaridade depreendidas a partir de definição para o 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*).

[1] Disponível em: http://www.nilc.icmc.usp.br/arianidf/sustento/

[2] O projeto SUSTENTO tem o potencial de subsidiar 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/

Visando alcançar os objetivos, as etapas metodológicas a seguir foram realizadas.

1.3 Metodologia

Equacionou-se metodologicamente esta pesquisa em 8 etapas ou tarefas apresentadas a seguir.

Tarefa 1 – Revisão da literatura: essa tarefa consistiu no estudo da teoria/modelo CST e dos fenômenos multidocumento de acordo com essa teoria, sobretudo a complementaridade. Ademais, a Tarefa 1 englobou a investigação dos métodos/atributos de identificação automática das relações CST ou similares.

Tarefa 2 – Seleção, recorte e estudo de *corpus*: a partir do *corpus* CSTNews, composto por coleções multidocumento de textos jornalísticos em PB anotados com base na teoria CST, essa etapa consistiu em recortar os pares de sentenças anotados com as relações de complementaridade (*Historical background*, *Follow-up* e *Elabaration*), gerando um subconjunto do CSTNews. Esse subconjunto foi divido em: *subcorpus* 1 e *subcorpus* 2. O *subcorpus* 1 foi usado para (i) estudo do fenômeno, que permitiu identificar suas características linguísticas e propor atributos que as traduzem, (ii) seleção dos atributos mais relevantes para distinguir as relações e os tipos de complementaridade e (iii) treinamento dos algoritmos de AM para gerar os classificadores. A unificação dos *subcorpora* 1 e 2 foi utilizada para teste e avaliação dos classificadores aprendidos a partir do *subcorpus* 1.

Tarefa 3 – Proposição e seleção de atributos: consistiu na tradução das características da complementaridade levantadas manualmente na Tarefa 2 em 9 atributos com potencial para discriminar automaticamente as relações CST e os tipos de complementaridade e na seleção dos atributos mais relevantes para tal tarefa. A relevância dos atributos foi analisada de forma manual e automática a partir do *subcorpus* 1. A análise manual consistiu na verificação da frequência de ocorrência dos atributos em função das relações e dos tipos. Na análise automática, a relevância foi calculada por um algoritmo de AM, que ranqueou os atributos em função do poder discriminativo de cada um deles.

Tarefa 4 – Estudo da correlação entre os atributos e as relações CST: consistiu na análise automática da correlação entre os atributos selecionados na Tarefa 3 e as relações CST de *Historical background, Follow-up* e *Elabaration*. A análise automática foi feita pela submissão dos *subcorpora* 1 e 2 unificados a algoritmos de AM da abordagem supervisionada.

Tarefa 5 – Estudo da correlação entre os atributos e os tipos de complementaridade: essa tarefa consistiu no estudo da correlação entre os atributos selecionados na Tarefa 3 e os tipos de complementaridade (temporais e atemporais). Essa correlação seguir as mesmas diretrizes da Tarefa 4.

1.4 Estrutura da dissertação

Esta dissertação está organizada em 6 capítulos. No Capítulo 2, apresenta-se a revisão da literatura. Especificamente, apresentam-se a teoria/modelo CST, destacando-se as relações de complementaridade, e as principais características linguísticas (ou atributos) utilizadas para a identificação automática das relações CST, incluindo as de complementaridade. No Capítulo 3, apresenta-se o recorte feito no *corpus* de referência CSTNews para a análise da complementaridade, resultando em um *subcorpus*, o qual foi divido em *subcorpus* 1 e *subcorpus* 2. Além disso, descreve-se o estudo manual da complementaridade no *subcorpus* 1, a partir do qual as características linguísticas desse fenômeno foram identificadas. No Capítulo 4, destaca-se como as características da complementaridade foram traduzidas em 9 atributos linguísticos com potencial de subsidiar a detecção automática das relações CST e dos tipos de complementaridade e como os mais relevantes dentre os 9 propostos foram identificados de forma manual e automática. No Capítulo 5, apresenta-se a análise automática, via AM, da pertinência dos atributos para identificar as relações CST e os tipos de complementaridade. Por fim, no Capítulo 6, tecem-se considerações finais sobre esta pesquisa, acompanhadas de apontamentos sobre a originalidade, trabalhos futuros e contribuições para a Linguística e o Processamento Automático das Línguas Naturais.

Capítulo 2

A CST E A IDENTIFICAÇÃO AUTOMÁTICA DA COMPLEMENTARIDADE

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

Inspirado na *Rhetorical Structure Theory*[3] (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 abordavam um 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 teórico multidocumento que estabelece 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 2.

Quadro 2: 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).

[3] O objetivo principal da RST é analisar um texto quanto a sua coerência. Para tanto, verifica se as unidades mínimas de discurso (*Elementary Discourse Units* - EDUs), que desempenham uma função para que o objetivo do texto seja atingido, estão conectadas entre si. Cada EDU é classificada em núcleo (informação principal) ou satélite (informação adicional). Quando coerente, um texto tem suas unidades conectadas entre si por meio de relações retóricas (também chamadas de relações de coerência ou discursivas), representadas na forma de árvore. Caso uma relação realize a conexão entre um núcleo e um satélite, tem-se uma relação mononuclear; caso conecte somente núcleos, tem-se uma relação multinuclear.

Radev (2000) aponta que os relacionamentos entre documentos que abordam um 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.

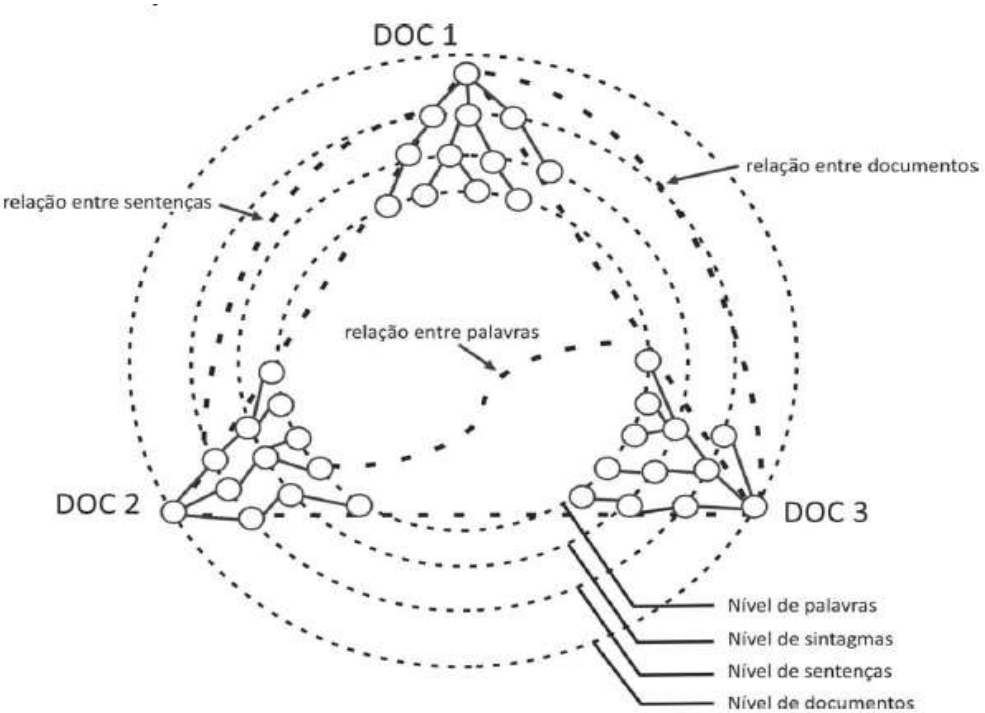

Figura 3: Esquema de relacionamento CST.
Fonte: Radev (2000).

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) estão representados por um subgrafo, que codifica relações internas aos textos. Os relacionamentos internos a cada texto podem ser estabelecidos em nível sintático ou semântico. 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 diferente; (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 3, 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 duas relações CST com direcionalidade.

Quadro 3:- 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, prevalecendo, então, a concordância entre os anotadores para a classificação dos pares de sentenças. Uma das alternativas encontradas por alguns autores para amenizar a ambiguidade existente entre os pares de sentenças foi revisar o conjunto original de relações proposto por Radev (2000).

Zhang *et al.* (2002) 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, unificaram relações do modelo original que consideraram similares, como as relações *Refinament*, *Description* e *Elaboration*, que foram unificadas a um único rótulo genérico, *Elaboration*. Além disso, os autores retiraram relações que não foram verificadas no *corpus*, como a relação *Change of perspective*, resultando em um conjunto de 14 rótulos.

A partir do refinamento de Aleixo e Pardo (2008), Maziero *et al.* (2010) elaboram uma tipologia para as 14 relações CST, ilustrada pela Figura 4.

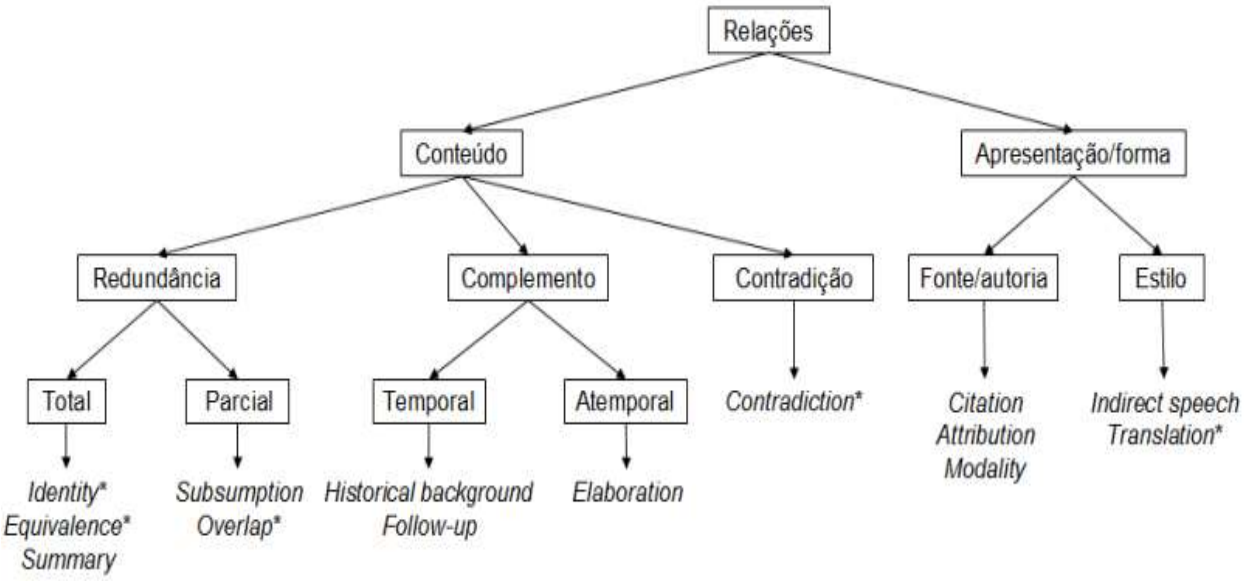

Figura 4: Tipologia das relações CST.

Fonte: Maziero *et al.* (2010).

Nessa tipologia, as relações CST foram organizadas em 2 grandes grupos: (i) relações de conteúdo, as quais rotulam os relacionamentos semânticos entre sentenças, e (ii) relações de forma, 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 3, são, respectivamente, de forma e de conteúdo (em especial, de redundância parcial).

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

Quadro 4: Definição das relações CST de Maziero *et al.* (2010).

Relação	Tipo	Direção	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: Adaptado de 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 no Capítulo 3.

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 as frequências das relações forem somadas em função das subcategorias de conteúdo (cf. Figura 4) tem-se a distribuição da redundância, complementaridade e contradição como ilustrada na Tabela 1.

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 for considerado 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, capturado pelas relações do modelo 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, definindo e exemplificando as relações CST que capturam esse fenômeno.

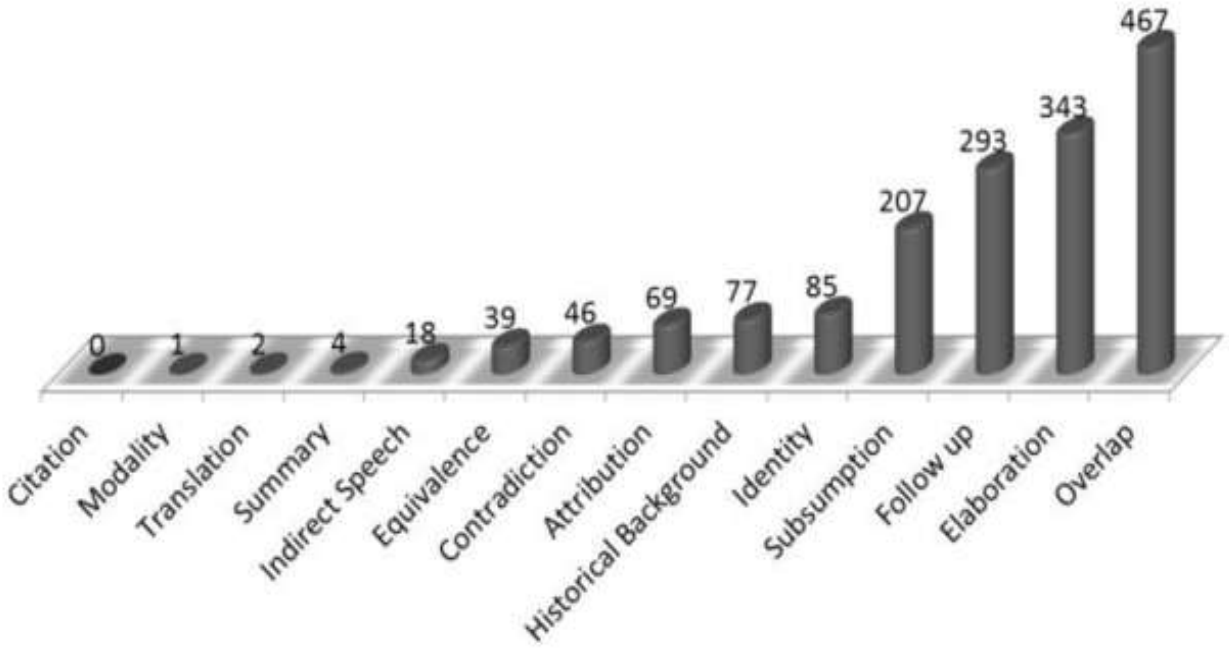

Figura 5: Frequência das relações CST no *corpus* CSTNews.

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	
Complementaridade	*Follow up*	293	713
	Historical background	77	
	Elaboration	343	
Contradição	*Contradiction*	46	46

Fonte: Adaptado de Maziero (2012).

2.2 As relações CST e a complementaridade

Na tipologia de Maziero *et al.* (2010), a complementaridade é uma subcategoria de relações de conteúdo. De acordo com os autores, a complementaridade é o segundo fenômeno multidocumento mais frequente no *corpus* CSTNews, perdendo apenas para a redundância.

De modo geral, entende-se que a complementaridade ocorre 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, uma das sentenças sempre possui informações adicionais em relação à outra. Em outras palavras, S1 e S2 possuem conteúdo em comum, sendo que S2 apresenta informação aditiva que não está presente em S1.

Ademais, segundo os autores, 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 5, retirados do *corpus* CSTNews, ilustram esses tipos de complementaridade.

Quadro 5: 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 5 por um par de sentenças provenientes de textos que relatam "um acidente aéreo Congo". Cada sentença é originária de textos distintos que abordam o mesmo assunto. 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 é capturado pela relação CST *Historical background*.

A complementaridade temporal do tipo (ii) é ilustrada por um par de sentenças que, advindas de textos distintos, possuem o mesmo tópico principal ("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

ao evento descrito em S1 após um intervalo curto de tempo. No caso, S2 fornece "o horário de abertura da pista (principal) para pouso", que ocorreu após a "abertura da pista auxiliar para decolagem" veiculado por S1. Segundo a tipologia apresentada por Maziero *et al.* (2010), esse tipo de complementaridade temporal é explicitado pela relação CST *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 5, 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 6, também retirados do *corpus* CSTNews, ilustram esse tipo de relação de conteúdo atemporal.

Quadro 6: Exemplos de complementaridade atemporal.

Complementaridade atemporal	**Sentenças**
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.

O primeiro par é formado por sentenças provenientes de textos que comunicam a "reforma da pista principal do aeroporto de Congonhas". Nele, observa-se que S1 e S2 possuem conteúdo comum ("cronograma da obra"), sendo que S2 fornece uma informação adicional sobre esse conteúdo. No caso, 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, 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, que é o foco de S2, diz respeito à "nacionalidade dos membros da tripulação" ("eram de nacionalidade russa"). De acordo com a Maziero *et al.* (2010), a complementaridade atemporal é codificada pela relação *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 semântias 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), a primeira etapa do método proposto consiste em analisar se há 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 (7). Caso o valor da *word overlap* obtido seja igual ou superior a 0.12[4], considera-se que as sentenças do par sob análise são relacionadas.

(7)

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

Em (7), vê-se que, para calcular a *word overlap* (*Wol*) entre um par de sentenças (S1 e S2) 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*[5], 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)[6], e (ii) distância semântica entre os núcleos de sintagmas nominais (SNs) e verbais (SVs) (atributo semântico). Para determinar a distância semântica entre as palavras nucleares em SNs e SVs, o método utiliza a WordNet de Princeton[7] (WN.Pr), uma base relacional de dados lexicais (FELLBAUM, 1998).

No caso do atributo morfossintático, quanto maior o número de etiquetas em comum entre as sentenças, maior a similaridade entre elas. No caso do atributo semântico, a

[4] 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.

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

[6] 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.

[7] 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.

similaridade é determinada pela proximidade da relação que 2 núcleos de SNs, por exemplo, possuem na hierarquia de conceitos da WN.Pr. Assim, 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 conjuntos de treinamento e teste compostos por 6 coleções de textos, cujas características estão na Tabela 2[8].

Tabela 2: *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).

Os pares de sentenças das coleções foram manualmente anotados com as relações CST. Os atributos necessários para a detecção da similaridade (isto é, *word overlap*) e de as relações CST (atributos lexical, sintático e semântico) também foram explicitados para cada par de sentenças. Além das 6 coleções da Tabela 2, os autores utilizam mais 1 coleção, denominada *Shuttle10* (cujo tópico é o acidente o *Space Shuttle Columbia*, em 2003), cujas sentenças não foram anotadas via CST, deixando explícitos somente os atributos.

Na sequência, as 7 coleções do *corpus* foram submetidas a algoritmos de AM que, a partir dos atributos explícitos, aprendem padrões estatisticamente relevantes e realizam 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: precisão[9], cobertura[10] e medida-F[11] (HIRSCHMAN; MANI, 2003). No

[8] Na Tabela 2, ressalta-se que o nome dado às coleções reflete a fonte da qual os textos da coleção foram coletados.

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

caso, o AM obteve os resultados descritos na Tabela 3, os quais incluem apenas as relações que tinham frequência maior que 20 nos dados de teste.

Tabela 3: Avaliação da identificação automática das relações CST de Zhang e Radev (2005)

Relação CST	Precisão	Cobertura	Medida-F
No relation	0.8875	0.9905	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 3, 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 se detectar.

Marsi e Krahmer (2005), por sua vez, não focalizam a identificação específica de relações CST, mas de relações semelhantes, a saber: *Equals, Generalizes, Specifies, Restates* e *Intersects*. Tais relações são definidas no Quadro 7, com base no exemplo de (8).

(8)

Sentença 1: *Dailly coffe diminishes risk on Alzheimer and Dementia.*

Sentença 2: *Three cups of coffee a day reduces chance on Parkinson and Dementia.*

O *corpus* utilizado consiste em duas traduções, em holandês, do livro "*Le petit prince*", escrito por Saint-Exupéry em 1943. De acordo com Marsi e Krahmer (2005), um *corpus* desse tipo garante quantidade considerável de sentenças relacionadas. Os cinco primeiros capítulos de cada tradução compõem o *corpus* de treinamento e o restante, o de teste.

Quadro 7: 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).

Para identificar as relações, as sentenças são representadas por árvores de dependência e os nós das árvores semanticamente correspondentes são alinhados, sendo o alinhamento rotulado pela relação do Quadro 7 correspondente. O alinhamento é semiautomático, baseando-se na (i) similaridade das dependências (*head/subject*, *head/modifier* e *coordination/conjunction*) e nas (ii) relações de sinonímia, hiperonímia e hiponímia, identificadas com base na EuroWordNet[12] (VOSSEN, 1998).

Uma vez alinhado e rotulado, o *corpus* de treinamento foi submetido a algoritmos de AM, que identificaram padrões estatisticamente relevantes para a detecção das relações *Equals, Generalizes, Specifies, Restates e Intersects*. Tais padrões foram aplicados ao *corpus* de teste para avaliar sua pertinência diante de um conjunto novo de dados. Tais resultados da avaliação estão expressos na Tabela 4.

Tabela 4: Avaliação da identificação automática das relações de Marsi e Krahmer (2005).

Relações	Precisão	Cobertura	Medida-F
Equals	0.99	0.97	0.98
Restates	0.65	0.82	0.73
Specifies	0.60	0.48	0.53
Generalizes	0.50	0.52	0.50
Intersects	0.69	0.35	0.46
Combinação	0.82	0.81	0.80

Fonte: Marsi e Krahmer (2005)

[12] A EuroWordNet é uma base multilíngue em que as bases construídas no formato da WN.Pr para várias línguas europeias, como o holandês, italiano, espanhol, alemão, francês, checo e estônio, estão interconectadas.

Com base na Tabela 4, os resultados das medidas de avaliação para identificação da relação *Equals* são maiores já que são baseados na identidade entre duas sentenças. A identificação das relações *Restates* e *Specifies* tem precisão similar. Isso parece ocorrer porque tais relações são bastante semelhantes entre si, caracterizando-se pela ocorrência de conceitos genéricos e específicos, respectivamente. A identificação de *Generalizes* tem a precisão mais baixa, já que capturar automaticamente a paráfrase é uma tarefa bastante custosa, evidenciado por autores como Seno e Nunes (2009). Por fim, a relação *Intersects* possui uma precisão relativamente boa, mas sua cobertura é bastante baixa. Possivelmente, esse resultado é obtido porque as sentenças do par não são idênticas e uma delas apresenta informações novas em forma de paráfrases, o que as aproxima das relações *Overlap* e *Elaboration* do modelo CST.

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**.

(9)

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

De acordo com a proposta de MacCartney *et al.* (2006) , em (9), a Sentença 1 expressa o texto, e a Sentença 2, a hipótese. Por meio de uma implicatura (ou acarretamento)[13] semântica, a Sentença 2 está compreendida na Sentença 1, já que, de fato, "mais de 2 milhões de pessoas" está compreendido por "2,5 a 3,5 milhões de pessoas".

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.

[13] A relação de acarretamento ocorre entre duas proposições (P1 e P2), em que a informação em P1 acarreta a informação expressa em P2. Por exemplo, "correr" acarreta "deslocar-se". Salienta-se que o acarretamento lexical é uma relação unilateral. No exemplo, "correr" acarreta "deslocar-se", entretanto o contrário não ocorre.

No método de MacCartney *et al.* (2006), *entailment* é determinado por um conjunto de 28 características ou atributos linguísticos, que 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 a adição de adjuntos sintáticos (adjuntos adverbiais, por exemplo), caracterizando-se por serem modificadores ou elementos que circunstanciam a ação na sentença. Por exemplo, "*Os cachorros latem*"[14] (em inglês, "*Dogs barked*") distingue-se de "*Os cachorros latem hoje*" (em inglês, "*Dogs barked today*"), já que "hoje", sintaticamente, caracteriza-se por ser um adjunto adverbial de tempo, circunstanciando a ação.

c) Antonímia: marcadores que evidenciam a polaridade entre um par de antônimos advindos do texto e da hipótese. Para tanto, os autores identificam os antônimos com base na WordNet.Pr e em uma lista de referência de antônimos.

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 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 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*") que são aplicados tanto ao texto, quanto à hipótese, e compara-se a distância entre os resultados.

[14] Tradução nossa.

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 em grafos, os autores mediram a precisão como métrica 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. Os autores apontaram que a acurácia máxima dessa tarefa foi de 0,65.

Miyabe *et al.* (2008) também investigaram a identificação automática de relações semelhantes às do modelo CST. Os autores identificaram, em especial, as relações *Equivalence* e *Transition*. De acordo com Miyabe *et al.* (2008), *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. No Quadro 8, exemplifica-se a ocorrência das relações *Equivalence* e *Transition*.

Quadro 8: Exemplo das relações *Equivalence* e *Transition*.

Texto 1

[1] *ABC said on the 18th that the number of users of its mobile-phone service had reached 1.500,000*. ("ABC disse, no dia 18, que o número de usuários de seu serviço de telefonia móvel tinha alcançado 1.500,000", tradução nossa)

[2] *Users can acces the internet, reserve train tickets, as well as make phone calls through this service*. ("Os usuários podem acessar a internet, fazer a reserva de passagens de trem, bem como fazer chamadas telefônicas por meio deste serviço", tradução nossa)

Texto 2

[1] *ABC telephone company announced on the 9th that the number of users of its mobile-phone service had reached one million*. ("A companhia telefônica ABC anunciou no dia 9 que o número de usuários de seu serviço de telefonia móvel tinha atingido um milhão", tradução nossa)

[2] *This service includes internet access, and enables train-ticket reservations and telephone calls.* ("Este serviço inclui acesso à internet, e permite reservar passagens de trem bilhetes e realizar chamadas telefônicas", tradução nossa)

Fonte: Miyabe *et al.* (2008)

De acordo com os autores, a Sentença 1 do Texto 1 e a Sentença 1 do Texto 2 presentes no Quadro 8 estabelecem 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), o que pode estar associado à variação de datas ("*on the 9th*" e "*on the 18th*", respectivamente). Já a Sentença 2 do Texto 1 e a Sentença 2 do Texto 2 estabelecem relação de *Equivalence*, pois as sentenças transmitem a mesma informação, ainda que na forma 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 possui 115 conjuntos de textos jornalísticos que abordam vários assuntos relacionados entre si. Os textos foram organizados em 15 coleções, em que cada uma delas possui, em média, 10 textos.

Para a identificação da relação *Equivalence*, os autores anotaram o *corpus* com as relações do modelo CST, e observaram que de, aproximadamente, 470.000 pares de sentenças, 798 possuíam tal relação.

Para a identificação da relação *Transition*, os autores propuseram um algoritmo[16], a saber: (i) identificar os sintagmas nominais (SNs) constituídos por valores numéricos, (ii) identificar os sintagmas em que os valores numéricos são dependentes em sintagmas preposicionais (SPprep), (iii) buscar os SNs que dependem dos SPpreps e (iv) extrair os SNs encontrados em (iii), exceto informações sobre data. No Quadro 8, "*one milion*" e "*1.500,000*" são valores numéricos, e "*of its mobile-phone servisse*" é o SP. Assim "*the number of users*" seria o SN em que os valores numéricos dependem.

A avaliação foi realizada com base nas medidas precisão, cobertura e medida-F. Para *Equivalence*, o método obteve 87,2 de precisão, 57,3 de cobertura e 69,2 de medida-F, enquanto que, para a detecção de *Transition*, o método obteve 27,4 de precisão, 41,2 de cobertura e 32,9 de medida-F.

[15] A medida cosseno é resultado de uma representação 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.

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

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, Mazeiro (2012) identifica as relações CST com base nas informações descritas no Quadro 9.

Vale ressaltar que, além das características sentenciais do Quadro 9, 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.

Quadro 9: Atributos para detecção automática das relações CST 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
12	Sobreposição de sinônimos entre S1 e S2

Fonte: Maziero (2012).

Para avaliar o CSTParser, era necessário utilizar um *corpus* anotado com as relações CST a fim de treiná-lo. Para tanto, foi construído o CSTNews (CARDOSO *et al.*, 2011). O conjunto de textos, então, foi dividido em duas parcelas: uma para treinamento e outra para avaliação. Na primeira parcela, um grupo de pesquisadores se reuniu por 3 meses para estudar as

relações CST e aplicá-las ao *corpus*. Ao término de cada treinamento, o grupo se reunia para analisar a concordância entre eles sobre as relações CST sugeridas para cada par de sentenças. Após isso, os anotadores se organizaram em três grupos e anotaram a segunda parcela do *corpus*, em que cada anotador deveria anotar todos os textos de cada coleção do *corpus*.

Ao avaliar o desempenho do método, Mazeiro (2012) obteve a precisão geral de 68,13%. Essa precisão geral é a média da precisão dos atributos do Quadro 9 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 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;* (ii) 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, 2004). O conjunto de teste inclui 100 pares compostos por sentenças sem anotação de relações CST.

[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.

[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.

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. *F-mesuare* > 90%) e na detecção dos pares sem relação CST (isto é, *F-mesuare* > 80%).

Segundo os autores, esses resultados podem ser decorrentes de dois fatores: as sentenças relacionadas por *Identity* apresentaram alta similaridade lexical e tamanho; em contrapartida, as sentenças sem relação CST não possuíam similaridade lexical e/ou de tamanho. 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 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 principiais métodos automáticos de identificação da redundância ou similaridade entre sentenças, enfatizando as 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 e John (2003) 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 estatístico e alguns métodos superficiais linguísticos foram analisados.

O método estatístico se baseia no número de palavras (de classe aberta) em comum entre as unidades de significado.

Os métodos superficiais linguísticos se baseiam na sobreposição de formas analisadas (canônicas) ou não-analisadas (formas que ocorrem na superfície textual); diz-se superficial-linguística, então, pelo fato de o método não necessitar de conhecimentos linguísticos de níveis profundos. Para calcular a sobreposição lexical, Hatzivassiloglou *et al.* (1999, 2001) utilizam a medida *word overlap.*

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 sofisticado 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 simples, Hatzivassiloglou *et al.* (1999, 2001) utilizam métodos compostos, os quais capturam dois tipos de característica das sentenças. Dentre eles, citam-se como exemplos:

[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/.

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

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

Além dos trabalhos de Hatzivassiloglou *et al.* (1999, 2001), Newman e John (2003) também focalizam métodos de detecção da redundância. Esses autores combinam o método superficial estatístico tradicional (ou seja, sobreposição de palavras) a um método linguístico, por meio do qual a similaridade é calculada com base em conhecimento de nível semântico. O método 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* (VOSSEN *et al.*, 2007) pela sinônima e/ou pela hiponímia. A aplicação desse

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

método parte da hipótese de que o compartilhamento de núcleos sintagmáticos semanticamente relacionados entre as sentenças de um par indica que estas são similares.

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 (7) (pág. 34). 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 (7) foi adaptada, gerando-se a fórmula *noun overlap* (Nol) descrita em (10). De forma análoga, fez-se a adaptação da medida original para o cálculo da sobreposição dos verbos, originando a fórmula *verb overlap* (Vol).

(10)

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

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.

[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.

Os atributos "sobreposição de núcleo de sujeito" e "sobreposição de núcleo de objeto/predicativo principal" também estão explicitamente citados na literatura. No entanto, eles podem ser vistos como especificações do atributo "sobreposição de núcleo de SN", já que buscam identificar não somente núcleos de SNs em comum, mas palavras que são núcleo em SNs com funções sintáticas específicas.

O atributo "sobreposição de palavra sinônima" foi selecionado por ser amplamente utilizados 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).

O atributo "sobreposição de etiqueta morfossintática" captura 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 de 45 pares de sentenças extraídas do *corpus* CSTNews. Esse conjunto se distribui como descrito na Tabela 5.

Tabela 5: Características do *corpus* de treinamento e teste de Souza *et al.* (2012).

Nível de redundância	Relação CST	Qt. de par por relação	Qt. de par por nível de redundância
Redundância total	*Identity*	5	15
	Equivalence	6	
	Summary	4	
Redundância parcial	*Subsumption*	8	16
	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) (WITTEN, FRANK, 2005) com o objetivo de investigar a adequação dos atributos 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 6, apresentam-se os testes ranqueados em função da precisão obtida para a distinção dos níveis de redundância (total, parcial ou nula).

Tabela 6: Teste automático dos atributos para a indicação dos níveis de redundância.

Teste	**Atributo**									**Precisão** (%)
	Loc	Wol	Nol	Vol	PdMorf	Suj	Vp	ObjPredp	MetMorf	
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).

O Teste 1 resultou no conjunto de regras evidenciado em (11), que obtiveram 97,7% de precisão, de acordo com o algoritmo PART do ambiente de aprendizado de máquina Weka. Dos 45 pares, apenas 1 foi classificado equivocadamente pela aplicação da regra 4.

(11)

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, 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 atributos para identificar as relações de redundância (*Identity, Equivalence, Summary*, *Subsumption* e *Overlap*) com precisão de 62.2%. Na Tabela 7, apresentam-se os resultados dos 11 testes realizados para apontar quais atributos melhor identificam as relações CST de redundância.

Tabela 7: Teste automático dos atributos para a indicação das relações CST de redundância.

Teste	**Atributo**									**Precisão** (%)
	Loc	Wol	Nol	Vol	PdMorf	Suj	Vp	ObjPredp	MetMorf	
1	■	■	■	■	■	■	■	■	■	62.2
11		■	■							60
3		■								55.5
4			■							51.3
5				■						48.8
6					■					44.4
7						■				42.2
8							■			42.2
9								■		42.2
10									■	42.2
2	■									13.3

Fonte: Souza *et al.* (2012).

No Teste 1, da Tabela 7, os autores utilizaram o algoritmo PART valendo-se dos 9 atributos de forma conjunta. Tal teste obteve precisão de 62.2%. Nos Testes 2 a 10, os autores testaram cada um dos atributos para identificar aqueles que são proeminentes em identificar as relações CST de redundância. Os Testes 3 e 4 testaram separadamente os atributos "*word overlap*" e "*noun overlap*", os quais obtiveram precisões 55.5% e 51.3%, respectivamente. Os autores, então, testaram conjuntamente os atributos que alcançaram as precisões mais altas. Assim, no Teste 11, utilizando somente os atributos "*word overlap*" e "*noun overlap*", é possível ter precisão de 60% em identificar as relações CST de redundância, aproximando-se da precisão obtida no Teste 1, o qual é baseado no uso dos atributos de forma conjunta.

2.5 Lições aprendidas

A revisão da literatura evidenciou que, na maioria dos trabalhos, as relações CST são automaticamente identificadas em função da similaridade entre as sentenças, posto que tais relações efetivamente estabelecem-se, em certa medida, entre sentenças que apresentam alguma sobreposição de conteúdo, e que, para a identificação da similaridade, há inúmeras estratégias disponíveis. Além disso, observou-se que a identificação das relações com base em regras, como ocorre, por exemplo, com *Contradiction* em Maziero (2012), requer o conhecimento prévio sobre as características específicas das relações. Assim, tendo em vista que a redundância é apenas uma das características da complementaridade enquanto fenômeno de conteúdo, investigou-se a complementaridade em *corpus* com o objetivo de identificar suas características específicas.

No Capítulo 3, apresenta-se o recorte feito no *corpus* de referência CSTNews para a análise da complementaridade. Esse recorte resultou em um subconjunto de pares de sentenças, o qual foi divido em *subcorpus* 1 e *subcorpus* 2. Além disso, descreve-se o estudo manual da complementaridade com base no *subcorpus* 1, a partir do qual as características linguísticas desse fenômeno foram identificadas.

Capítulo 3

SELEÇÃO, RECORTE E ESTUDO DE *CORPUS*

3.1 O *corpus* CSTNews

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

O CSTNews está organizado em 50 *clusters* (ou coleções). Cada coleção aborda um assunto distinto, sendo cada texto (notícia) da coleção proveniente de um jornal distinto. No total, o CSTNews 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", "dinheiro" e "esporte". No Quadro 10, tem-se a distribuição dos *clusters* em função de sua categoria.

Quadro 10: 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
Dinheiro	C30

Fonte: Elaborado pelo autor.

Cada *cluster* é constituído por: (i) 2 ou 3 textos-fonte, (ii) sumário manual de cada texto-fonte, (iii) sumário manual multidocumento, (iv) sumário automático multidocumento, (v) 5

abstracts e 5 extratos elaborados manualmente, (vi) interconexão entre os textos-fonte via CST, (vii) anotação de expressões temporais nos textos-fonte, (viii) etiquetação morfossintática e sintática dos textos-fonte, (ix) anotação dos sentidos dos substantivos e verbos via *synsets* da WN.Pr, (x) anotação de aspectos informacionais de um dos sumários multidocumento de referência (manuais) (p.ex.: *o quê*, *onde*, etc.), (xi) anotação discursiva de cada texto-fonte via RST, e (xii) anotação de subtópicos informativos dos texto-fonte.

Todos os sumários do CSTNews possuem taxa de compressão de 70%, ou seja, apresentam 30% do número de palavras de seus respectivos textos-fonte. Para os sumários multidocumento, a taxa de compressão de 70% é calculada a partir do maior texto do *cluster*.

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 (ALEIXO, PARDO, 2008) que, dada 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.

Na Tabela 8, mostra-se a porcentagem de cada relação no CSTNews.

Tabela 8: Frequência de ocorrência das relações no CSTNews.

Relação CST	**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).

Sobre a Tabela 8, 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, apesar de isso não caracterizar um impedimento para a ocorrência do fenômeno (cf. Quadro 4).

3.2 O *subcorpus* de complementaridade

Para descrever especificamente a complementaridade, fez-se um recorte no CSTNews, que consistiu em selecionar, por meio da interface *online*[23] de consulta ao *corpus*, apenas os pares de sentenças anotadas com as relações CST do tipo complementaridade, ou seja, *Follow-Up, Historical Background* e *Elaboration*. Esse recorte resultou em um subconjunto do CSTNews, cujos dados quantitativos estão descritos na Tabela 9.

Tabela 9: A complementaridade no *corpus* CSTNews.

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

Fonte: Elaborado pelo autor.

Com base na Tabela 9, observa-se que a complementaridade atemporal, decodificada pela relação *Elaboration*, ocorre em 343 pares de sentenças. Já a complementaridade temporal ocorre em 370 pares de sentenças, sendo 293 ocorrências da relação *Follow-up* e 77 ocorrências da *Historical background*. Assim, o subconjunto é composto por 713 pares de sentenças. Na Figura 6, apresenta-se a distribuição percentual dos pares do *subcorpus*.

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

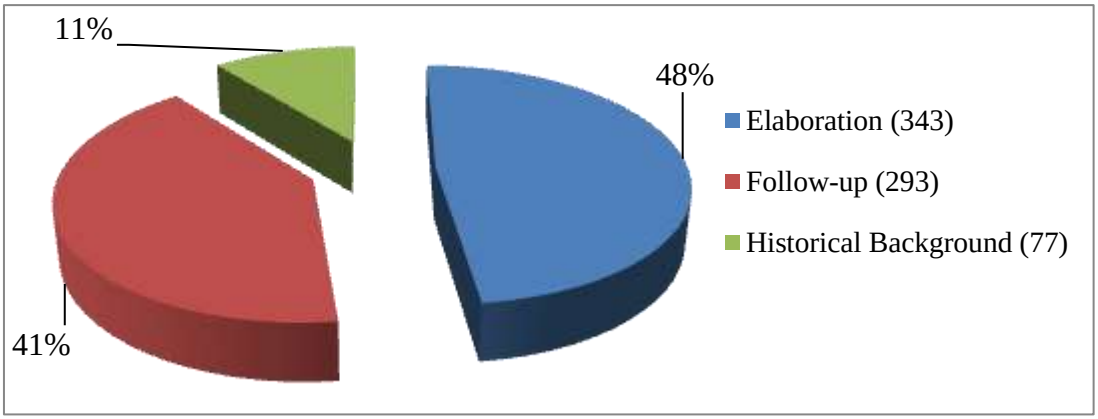

Figura 6: Distribuição percentual das relações de complementaridade no CSTNews.
Fonte: Elaborado pelo autor.

3.3 Os *subcorpora*

Diante dos objetivos de caracterizar o fenômeno da complementaridade, definindo atributos que subsidiam a detecção automática das relações CST dos diferentes tipos de complementaridade, e de testar os atributos de forma automática, o referido subconjunto de pares de sentenças do CSTNews foi dividido em *subcorpus* 1 e *subcorpus* 2.

O *subcorpus* 1 é composto por 135 pares de sentenças dos 713 anotados com as relações de complementaridade (cf. Figura 6), sendo: (i) 45 pares de sentenças anotadas com a relação *Elaboration* (isto é, 13,11% do total de 343), (ii) 45 pares de *Historical background* (o que representa 58,44% do total de 77) e (iii) 45 de *Follow-up* (isto é, 15,35% do total de 293).

Quanto ao *subcorpus* 1, vale ressaltar que, ao se considerar o tipo de complementaridade (temporal e atemporal), observa-se um desbalanceado dos dados, pois há 45 pares anotados com a relação atemporal *Elaboration* e 90 pares de sentenças anotados com relações de complementaridade temporal, isto é, *Historical background* e *Follow-up*. Esse desbalanceamento, que teoricamente privilegiaria a descrição da complementaridade temporal em detrimento da atemporal, não foi problemático, pois, ao se analisar outros pares do *subcorpus* inicial (distintos dos que compõem o *subcorpus* 1), observou-se que as características relativas ao tipo temporal até então identificadas no *subcorpus* 1 se repetiam, não havendo características relevantes que já não tivessem sido mapeadas no *subcorpus* 1. Em outras palavras, o *subcorpus* 1 se mostrou suficiente para identificar as características dos diferentes tipos de complementaridade, apesar o desbalanceamento.

O *subcorpus* 2 é composto por pares de sentenças distintos dos que compõem o *subcorpus* 1. Especificamente, o *subcorpus* 2 engloba 20% de cada relação que codifica a complementaridade, a saber: (i) 69 pares de sentenças anotadas com *Elaboration* (isto é, 20%

do total de 343), (ii) 61 pares com relação *Follow-up* (ou seja, 20% de 293), e (iii) 17 de *Historical background* (isto é, 20% de 77).

Na Tabela 10, tem-se a distribuição dos dados nos *subcorpora*.

Tabela 10: A distribuição dos dados nos *subcorpora*.

Complementaridade	**Relação CST**	***Subcorpus* 1**	***Subcorpus* 2**
Atemporal	*Elaboration*	45	69
Temporal	*Follow-up*	45	61
	Historical background	45	17
Total	--	135	147

Fonte: Elaborado pelo autor.

O *subcorpus* 1 foi destinado à descrição ou estudo (manual) do fenômeno da complementaridade. Essa descrição resultou em um conjunto de características linguísticas, que foram traduzidas em atributos para subsidiar a detecção automática dos tipos e relações CST de complementaridade. Ademais, esse *subcorpus* serviu de base para o processo de "seleção de atributo", que originou um subconjunto atributos, composto apenas pelos mais relevantes para a detecção das relações e tipos de complementaridade. Por fim, salienta-se que o *subcorpus* 1 foi utilizado para o treinamento dos algoritmos de AM. Por esse motivo, pode-se dizer que o *subcorpus* 1 é o "*corpus* de treinamento" deste trabalho.

Para testar e avaliar os padrões estatisticamente relevantes aprendidos pelos algoritmos de AM (os quais subsidiam os chamados classificadores) somente com base no *subcorpus* 1, o *subcorpus* 1 foi unificado ao *subcorpus* 2. Dessa maneira, os atributos foram automaticamente testados/avaliados em um único conjunto de dados, composto por 282 pares de sentenças resultante da unificação dos *subcorpora* 1 e 2 (135+147), a saber: (i) 114 pares de sentenças com a relação *Elaboration* (ou seja, 45+69), de um total de 343; (ii) 106 de *Follow-up* (ou seja, 45+61), de um total de 293, e (iii) 62 pares anotados com *Historical background* (isto é, 45+17), de um total de 77.

A seguir, descreve-se a análise manual da complementaridade no *subcorpus* 1.

3.4 A descrição manual da complementaridade

3.4.1. Características gerais da complementaridade: redundância

Como observado por Zhang e Radev (2005), as relações CST sempre ocorrem entre sentenças que são semanticamente relacionadas. Assim, elas ocorrem entre sentenças que possuem conteúdo em comum, em menor ou maior grau, dependendo da relação CST em questão. Essa característica das relações CST, aliás, justifica o fato de que a maioria dos métodos automáticos de identificação das relações CST é baseada na identificação da similaridade ou redundância entre as sentenças. Por esse viés, quanto à organização da tipologia proposta por Maziero *et al.* (2010), pode-se organizar hierarquicamente os fenômenos multidocumento quanto ao grau de redundância, em que redundância > complementaridade > contradição. Quanto à complementaridade, a própria definição das relações CST que codificam esse fenômeno evidencia tal característica. A similaridade de conteúdo entre sentenças de um par pode se expressar por meio de vários aspectos ou características como: (i) material lexical, (ii) localização no texto-fonte e (ii) subtópico.

a) Similaridade lexical

Sabe-se que o material lexical (ou seja, as palavras) que constitui as sentenças, em especial as palavras de classe aberta, pode relevar o grau ou nível de similaridade entre elas. Trabalhos como os de Hatzivassiloglou *et al.* (1999, 2001), Newman e John (2003) e Hendrickx *et al.* (2009), para o inglês, e Souza *et al.* (2012), para o português, evidenciam o potencial do material lexical para a caracterização da redundância ou similaridade. Em (10), há um par de sentenças complementares do *subcorpus* 1, anotadas com a relação *Elaboration.* Nelas, é possível observar a similaridade em função do material lexical.

(10)

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 (10) veiculam informação sobre uma pesquisa de intenção de voto. Ambas veiculam a informação sobre o período em que a pesquisa foi realizada, o que pode ser visto pela ocorrência em comum das palavras (de classe aberta): “pesquisa”, “dia” e “julho”.

b) A localização no texto-fonte

De acordo com Souza *et al.* (2012), a localização das sentenças de um par em seus respectivos texto-fonte também é um aspecto linguístico que salienta o grau de 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).

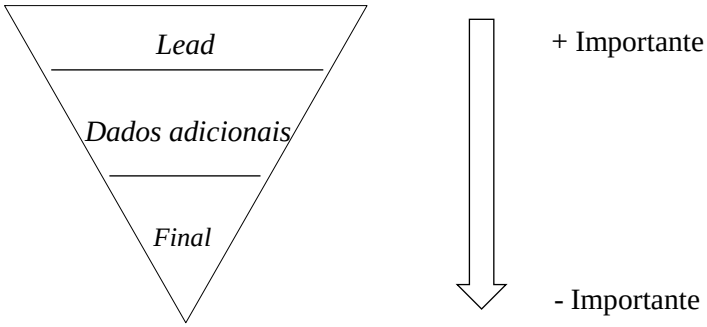

Figura 7: Estrutura do texto jornalístico: Pirâmide invertida.
Fonte: Adaptado de Lage (2002).

O método da pirâmide invertida ordena a informação de forma decrescente de relevância, sendo o texto 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 Sentença 1 do Texto 1 e a Sentença 1 do Texto 2 do Quadro 11[24] estão conectadas pela 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) quantas pessoas ficaram feridas (“25 ficaram feridas”), (iii) o dia (“segunda-feira”) e (iv) o local do incidente (“Moscou”).

[24] O traço pontilhado no Quadro 11 representa a distinção de subtópicos, a ser explicada na subseção seguinte.

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 11: Exemplos de redundância em função da localização no texto-fonte.

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.
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/C15

No caso de pares de sentenças complementares, a distância entre elas parece ser maior, pois, ao compartilharem menos conteúdo que as sentenças conectadas por relações de redundância, as sentenças tendem a ocupar posições mais distantes nos textos-fontes. Por exemplo, entre a Sentença 2 do Texto 1 e a Sentença 7 do Texto 2, 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 Texto 2 acrescenta a informação de que o governo russo descarta a possibilidade de a explosão ser um atentado terrorista.

c) O subtópico textual

Um texto jornalístico veicula um assunto principal (tópico), que é expresso no *lead*, e detalhes sobre ele (subtópicos), os quais são expressos no corpo do texto por segmentos que podem se ligar direta ou indiretamente ao tópico de acordo a progressão temática (cf. KOCH, 2009). Assim, se a similaridade de localização das sentenças nos textos-fonte evidencia similaridade de conteúdo (cf. SOUZA *et al.*, 2012), o mesmo acontece com a similaridade de subtópicos.

No caso, a sobreposição de subtópicos indica maior similaridade e a ocorrência de subtópicos distintos entre elas pode indica menor similaridade.

Segundo a anotação de subtópicos do CSTNews realizada por Cardoso *et al.* (2012), os Textos 1 e 2 do Quadro 11 estão segmentados em subtópicos, o que é indicado pelas linhas tracejadas que separam as sentenças. De acordo com os autores, as Sentenças 1, 2, 3 e 4 do Texto 1 estão cobertas pelo subtópico "a explosão", e as Sentenças 5 e 6 pelo subtópico "o motivo da explosão". Já as Sentenças 1, 2, 3, 4, e 5 do Texto 2 possuem o subtópico "a explosão", e as Sentenças 6 e 7 "intervenções do governo". Considerando-se a Sentença 1 do Texto 1 e a Sentença 1 do Texto 2, que foram anotadas com a relação CST de redundância *Equivalence*, observa-se que ambas possuem o mesmo subtópico ("a explosão").

Além dessas características gerais, que se pautam basicamente na similaridade (ou sobreposição) de conteúdo, a análise manual do *subcorpus* 1 permitiu identificar características específicas dos diferentes tipos de complementaridade (temporal e atemporal), as quais são descritas na sequência.

3.4.2. Características específicas da complementaridade temporal

a) Os advérbios

Nos pares de sentenças rotulados com as relações CST *Historical background* e *Follow up,* a informação complementar entre elas está diretamente ligada a questões temporais. Com base em alguns exemplos do *subcorpus* 1, observou-se que essa complementaridade se expressava pela ocorrência de advérbios de tempo. Em (11), ilustra-se essa característica com um par cujas sentenças estão conectadas pela relação *Follow-up*.

(11)

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.

No exemplo, a informação principal compartilhada pelas sentenças é a de que houve conflito entre o Hezbollah (grupo terrorista) e (as forças armadas de) Israel no domingo. Essa informação em comum está expressa no trecho "ataques do Hezbollah no domingo" da Sentença 1 e em toda a Sentença 2. A informação complementar à principal, de que "Israel lançou uma ofensiva depois dos ataques do Hezbollah (no domingo)", está expressa na Sentença 1 e só é reconhecida como posterior à principal (ou seja, *Follow-up*) por causa da ocorrência do advérbio "depois". Assim, a presença de advérbios de tempo nas sentenças que expressam a complementaridade temporal é uma característica desse tipo de complemento.

b) As expressões de tempo

Além dos advérbios, existem expressões em português que também indicam tempo. De acordo com Baptista *et al.* (2008), as expressões temporais (ETs) se organizam em 4 grupos, como exemplificado na Figura 8.

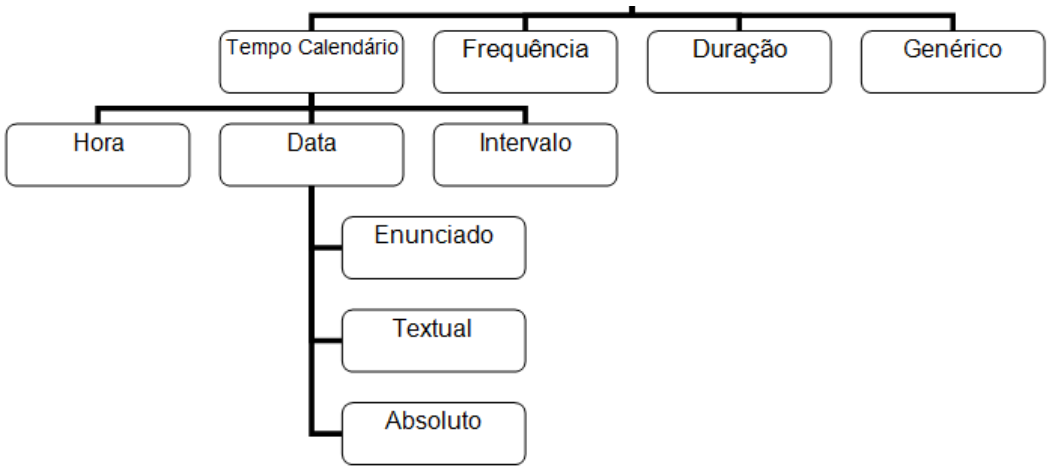

Figura 8: Tipologia das Expressões Temporais.
Fonte: Adaptado de Menezes Filho e Pardo (2011)

Segundo a organização proposta por Menezes Filho e Pardo (2011), demonstrada na Figura 8, as 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", em especial, 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 as ETs do subtipo "data" podem ser do tipo: (i) enunciado (p.ex.: "Partiu em março"), (ii) textual (p.ex.: "Um dia após a venda") ou (iii) absoluto (p.ex.: "O acidente ocorreu em fevereiro de 2002").

Menezes Filho e Pardo (2011) utilizaram a tipologia de Baptista *et al.* (2008) para anotar as ETs no *corpus* CSTNews. Especificamente, eles identificaram aproximadamente 1.000 ETs em todo o *corpus*. Em (12), há exemplos de expressões em um par de sentenças complementares que foi anotado com a relação *Historical background*.

(12)

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 (12), as sentenças compartilham a informação de que há acidentes aéreos no Congo. Na Sentença 2, tem-se que a informação principal é um acidente específico que ocorreu na quinta-feira, matando 17 pessoas. Nessa sentença, ocorre a expressão "na quinta-feira" que é do subtipo "data", de acordo com a referida tipologia. Na Sentença 1, tem-se a informação complementar à principal, que é o fato de que os acidentes aéreos são frequentes na região do Congo. Nessa sentença, a expressão "são frequentes" foi anotada como uma ET do subtipo "frequência" no CSTNews.

Assim, a ocorrência de advérbios de tempo e de ET nas sentenças em que há a informação complementar, é uma evidência linguística que caracteriza a complementaridade temporal.

3.4.3. Características específicas da complementaridade atemporal

A complementaridade atemporal parece não possuir muitos traços ou marcas linguísticas na superfície textual como a complementaridade temporal. No entanto, alguns marcadores discursivos podem indicar a ocorrência de detalhamento ou informação adicional.

Taboada e Das (2013) apontam que há teorias que estudam relações que se estabelecem de forma intersentenciais e/ou intertextuais com base em análises de sequências discursivas. Os modelos RST e CST são teorias que se estabelecem dessa maneira. O relacionamento entre sequências discursivas pode ser identificado, por vezes, com base em marcas linguísticas que se materializam na superfície textual.

O analisador discursivo automático Dizer 2.0, por exemplo, construído por Maziero e Pardo (2010) para o português, é uma ferramenta de PLN que identifica as relações que se estabelecem entre proposições de um mesmo texto de acordo com o modelo RST. E essas relações são automaticamente detectadas pela ocorrência de alguns marcadores discursivos (Quadro 12).

Quadro 12: Marcadores discursivos de complementaridade do Dizer 2.0.

Marcador	**Relação RST**
adicionalmente	Atemporal
ainda*****	Atemporal
além de*****	Atemporal
além disso	Atemporal
analogamente	Atemporal
após	Temporal
assim	Temporal
atualmente	Temporal

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

Fonte: Elaborado pelo autor.

(13)

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 <u>ainda assim</u> a aeronave da TAM, um Airbus A320, continuou voando, com o reverso direito desligado.

Em (13), as sentenças do par informam sobre um acidente aéreo no Brasil. A Sentença 2 traz informações complementares que, por sua vez, não estão contidas na Sentença 1, como o

modelo da aeronave (“Airbus A320”) e informações do acidente (“continuou voando, com o reverso direito desligado”). Essas informações são introduzidas pelo marcador discursivo “ainda assim”.

3.4.4. A complementaridade linguisticamente não-marcada

Em (14), (15), (16) e (17), 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.

(14)

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 (14), não há marcas linguísticas evidentes que se relacionam à temporalidade ou à atemporalidade. Ainda assim, é possível perceber a complementaridade. No caso, a informação complementar na Sentença 2 é o “critério técnico”, que será aplicado aos “recursos” mencionados na Sentença 1. Dessa forma, a percepção da complementaridade decorre principalmente de inferência baseada em conhecimento de mundo (p.ex.: Lula é presidente e, por isso, tem o poder para destinar recursos).

(15)

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 (15), não existem marcas linguísticas que explicitamente evidenciam a complementaridade. 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). Para tanto, o falante necessita resolver as anáforas existentes, de forma a identificar a informação complementar, em (15), que é o motivo do decreto do estado de alerta (no caso, “a forte chuva em São Paulo”).

(16)

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 (16), 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. Na Sentença 1 o falante tenta resolver a correferência para compreensão do conteúdo, permitindo-lhe 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”. Além disso, pode haver uma relação de parte-todo ou acarretamento entre os termos destacados, podendo evidenciar a informação complementar.

(17)

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 (17), 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 lhe permite inferir que “a próxima fase” de S2 é a “fase final da competição” veiculada por S1.

Com base nas características da complementaridade aqui descritas, apresenta-se, no próximo Capítulo, (i) a tradução das mesmas em 9 atributos capazes de subsidiar a tarefa automática de detectar as relações CST e os tipos de complementaridade e a (ii) seleção dos atributos mais relevantes dentre os 9 inicialmente propostos, o que foi feito por meio de uma análise manual e outra automática (via AM).

Capítulo 4

PROPOSIÇÃO E SELEÇÃO DE ATRIBUTOS

4.1 Delimitação dos atributos que tipificam a complementaridade

As características do Capítulo 4 foram traduzidas em 9 atributos, conforme o Quadro 13.

Quadro 13: Atributos para a caracterização da complementaridade.

Fenômeno	Atributo	Descrição
Complementaridade	Sobreposição de nome	Captura a redundância com base na sobreposição de nomes entre as sentenças de um par.
	Distância	Captura a redundância com base na localização das sentenças de um par.
	Ocorrência de advérbio na Sentença 1	Identifica a complementaridade temporal com base na ocorrência de advérbios temporais na primeira sentença de um par, desde que estejam relacionadas à complementaridade.
	Ocorrência de advérbio na Sentença 2	Identifica a complementaridade temporal com base na ocorrência de advérbios temporais na segunda sentença de um par, desde que estejam relacionadas à complementaridade.
	Ocorrência de expressão temporal na Sentença 1	Captura a complementaridade temporal com base na ocorrência de ETs na primeira sentença de um par, desde que estejam relacionadas à complementaridade.
	Ocorrência de expressão temporal na Sentença 2	Captura a complementaridade temporal com base na ocorrência de ETs na segunda sentença de um par, desde que estejam relacionadas à complementaridade.
	Sobreposição de subtópico	Captura a redundância com base na sobreposição de subtópicos entre os pares de sentenças.
	Ocorrência de marcador discursivo na Sentença 1	Identifica a complementaridade atemporal com base na ocorrência de marcadores discursivos na primeira sentença de um par, desde que estejam relacionadas à complementaridade.
	Ocorrência de marcador discursivo na Sentença 2	Identifica a atemporal com base na ocorrência de marcadores discursivos na segunda sentença de um par, desde que estejam relacionadas à complementaridade.

Fonte: Elaborado pelo autor.

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

A "sobreposição de nome", como mencionado, é um dos atributos mais relevantes para a identificação da redundância. Para especificar esse atributo entre as sentenças de um par, selecionou-se a medida *noun overlap* (Nol) utilizada por Souza *et al.* (2012) e que está descrita em (10).

O atributo "distância" também foi definido segundo as diretrizes de Souza *et al.* (2012). Assim, quanto menor a distância entre as posições das sentenças nos textos-fonte, maior a similaridade entre elas. Caso contrário, quanto maior a distância, menor a sobreposição de conteúdo. Segundo essa definição, o par em (18), composto por sentenças conectadas pela relação CST *Historical background* que ocupam a mesma posição em seus respectivos textos-fonte, tem o atributo "distância" com valor 0, revelando alta similaridade entre elas.

(18)

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.

O atributo "ocorrência de advérbio" busca codificar se as sentenças de um par possuem advérbios temporais envolvidos diretamente na complementaridade. Esse atributo foi especificado em função da ocorrência do advérbio na Sentença 1 e/ou na Sentença 2 do par. Assim, passou-a a ter dois atributos específicos, "advérbio em S1" e "advérbio em S2".

Em (19), as sentenças do par relatam sobre conflitos militares em regiões israelenses. O par de sentenças foi anotado com a relação *Historical background*. Na Sentença 1, há o advérbio "hoje", denotando o aspecto temporal na sentença. Na Sentença 2, há a informação complementar (trecho sublinhado), em que as "fontes militares" apontaram a quantidade de plataformas de lançamento de foguetes que tinham sido destruídas até aquele momento

(19)

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.

O atributo "ocorrência de expressões temporais" busca capturar se as sentenças de um par possuem ETs envolvidas diretamente na complementaridade. Esse atributo também 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 em S1" e "ET em S2".

No exemplo em (20), o par de sentenças descreve um acidente aéreo no Brasil, com uma aeronave da empresa aérea TAM. As sentenças foram anotadas com a relação *Historical background*. Na Sentença 2, a ET "desde o último dia 13" evidencia o aspecto temporal. A informação história ocorre na Sentença 1, por meio da construção "em 1996".

(20)

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.

O atributo "sobreposição de subtópico" busca identificar se as sentenças do par possuem, ou não, o mesmo subtópico textual. A hipótese para a proposição desse atributo é a de que sentenças complementares tendem a apresentar subtópicos distintos, já que a redundância entre elas é baixa (ou seja, as sentenças compartilham pouca informação entre si).

Em (21), o par de sentenças foi anotado com a relação *Follow-up*. De acordo com a anotação realizada por 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 Hezbollah". Assim, as sentenças não mantêm sobreposição de subtópico.

(21)

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 "ocorrência 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: "marcador discursivo em S1 e em 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).

Em (22), as sentenças do par foram anotadas com a relação *Elaboration*. As sentenças informam sobre um acidente aéreo no Brasil. A informação complementar ("a aeronave da TAM, um Airbus A320, continuou voando, com o reverso direito desligado"), na Sentença 2, é introduzida pelo um marcador discursivo "ainda assim".

(22)

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.

4.2 Seleção de atributos

A partir da proposição dos 9 atributos, realizou-se a análise manual e automática da relevância dos atributos quanto à discriminação dos tipos e relações CST de complementaridade com o objetivo de selecionar os mais pertinentes, reduzindo, assim, o conjunto inicial.

A "seleção de atributos" é importante por várias razões. Uma delas diz respeito ao fato de que os algoritmos de AM tendem a lidar mais adequadamente com um número menor de atributos, aprendendo com mais eficiência as classes e, consequentemente, classificando mais

adequadamente as instâncias. Outra razão se relaciona ao baixo custo computacional da detecção da complementaridade, pois, a aplicação de um conjunto reduzido de atributos reduz também o volume de informação linguística a ser explicitada no *corpus* para que a identificação possa ser feita.

A efetiva seleção dos atributos foi feita de forma manual e automática, além de considerar-se os tipos e as relações CST de complementaridade. A seleção manual consistiu em identificar os atributos mais pertinentes com base na frequência de ocorrência dos mesmos no *subcorpus* 1 quanto aos tipos e relações. A análise automática consistiu em aplicar um algoritmo de AM que ranqueou os atributos em função do poder discriminativo de cada um deles.

Para essa análise, fez-se um pré-processamento do *subcorpus* 1, que consistiu em explicitar ou descrever de forma semiautomática todas as informações linguísticas das sentenças dos pares necessárias ao cálculo ou verificação dos atributos entre elas.

Para o cálculo da "sobreposição de nome" e "ocorrência de advérbio", foi preciso explicitar todas as palavras pertencentes a essas classes que compunham as sentenças *subcorpus* 1. Para tanto, ao *subcorpus* 1, aplicou-se o PALAVRAS (BICK, 2000), um analisador sintático automático (ou *parser*) que realiza a tarefa de etiquetação morfossintática, ou seja, realiza a associação de uma única etiqueta às palavras de um texto que codifica sua correta classe gramatical. Assim, identificou-se automaticamente a classe de todas as palavras das sentenças do *subcorpus* 1, permitindo a identificação e classificação mais eficiente dos nomes e advérbios.

O cálculo da "distância" entre as sentenças, por sua vez, requereu as respectivas localizações em seus textos-fonte, que foi recuperada manualmente do CSTNews por meio de sua interface *online*.

Para a verificação da "ocorrência de expressões temporais", explicitaram-se as ETs em cada uma das sentenças 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*.

Para verificar a "sobreposição de subtópico", os subtópicos de cada sentença também foram manualmente recuperados de uma anotação prévia do CSTNews, no caso, realizada por Cardoso *et al.* (2012), e disponibilizada na página do projeto SUCINTO[25].

[25] http://www.icmc.usp.br/~taspardo/sucinto/cstnews.html

Por fim, o atributo "ocorrência de marcador discursivo" requereu a verificação manual da ocorrência dos marcadores nas Sentenças 1 e 2 de cada par, o que foi feito com base na lista de marcadores discursivos de Maziero e Pardo (2011), utilizada no Dizer 2.0.

No Quadro 14, ilustra-se a descrição das informações linguísticas subjacentes a cada um dos atributos, a qual foi organizada em uma tabela no formato *xlsx*. Para facilitar a visualização dos dados, os títulos das colunas, que indicam as informações subjacentes aos atributos, foram abreviados no Quadro 14: Dist (distância), N (nome), Adv (advérbio), ET (Expressão Temporal), SubT (subtópico) e MD (marcador discursivo). Ademais, cada linha equivale a uma sentença do *subcorpus.* Nas colunas, registram-se: (i) identificação numérica do par, (ii) *cluster*, (iii) relação CST, (iv) numeração da sentença, e (iv) descrição das informações linguísticas subjacentes aos atributos.

Quadro 14: Exemplo das informações linguísticas subjacentes aos atributos.

Corpus				***Descrição Linguística***					
Par	***Cluster***	**Relação CST**	**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	data	1	nsa
57	10	*Historical Background*	S1	6	fonte, miliciano, Hisbolá, foguete, Katyusha, mísseil, alcance, território, início, conflito	hoje	nsa	3	nsa
			S2	6	Comando, guerrilheiro, libanês, cidade, Tiro, plataforma, lançamento, foguete, fonte	nsa	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	data	1	nsa

Fonte: Elaborado pelo autor.

Quanto a S1 do par 6, por exemplo, observa-se que a tarefa de caracterização das sentenças gerou as seguintes informações linguísticas:

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.

Após a descrição ilustrada pelo Quadro 14, procedeu-se à verificação ou cálculo manual dos atributos entre as sentenças dos pares, cujos resultados foram analisados de forma manual e automática com o objetivo de eleger os atributos mais discriminativos.

O atributo "sobreposição de nome" foi calculado estatisticamente, gerando, assim, valores numéricos entre 0 e 1. Para tanto, optou-se pela medida *noun overlap* utilizada por Souza *et al.* (2012). Aplicando essa medida, o par 6 do Quadro 14 possui "sobreposição de nome" com valor 0,05, indicando pouco conteúdo em comum. Esse valor resulta do fato de que, entre os 10 nomes distintos presentes na Sentença 1 (porta-voz, avião, Soviet, Antonov-28, fabricação, propriedade, companhia, Trasept Congo, carga, mineral) e os 9 nomes constitutivos da Sentença 2 (acidente, localidade, Bukavu, leste, República Democrática do Congo, RDC, pessoa, porta-voz, Nações Unidas), somente "porta-voz" é comum.

O cálculo do atributo "distância" também foi baseado em Souza *et al.* (2012). Considerando o par 6, pertencente ao do *cluster* 1 (cf. Quadro 14), a Sentença 1 e a Sentença 2 ocorrem nas posições 6 e 1 em seus textos-fonte, respectivamente. Assim, a distância entre elas é 5. Sobre esse atributo, vale ressaltar que o resultado do cálculo obtido para cada par foi normalizado, devido ao fato de que os textos têm tamanhos diferentes. O resultado do cálculo do atributo "distância" foi normalizado em função da maior distância observada entre sentenças do respectivo *cluster*. Para tanto, verificou-se manualmente a maior distância para cada conjunto de textos do CSTNews. Por exemplo, o par 6 faz parte do *cluster* 1, o qual tem 6 (seis) como a maior distância entre os pares de sentenças. Assim, todos os valores do atributo "distância" obtidos para pares desse *cluster* foram divididos por esse valor. Dessa maneira, a distância normalizada do par 6 é aproximadamente 0.83 (isto é, 5/6).

Para verificar os atributos "ocorrência de advérbio em S1" e "ocorrência de advérbio em S2", verificou-se, com base na descrição das sentenças, se havia sido assinalada a ocorrência

de advérbios de tempo em cada uma das sentenças de um par, quando esses estivessem diretamente relacionados à complementaridade. Dessa forma, com relação ao par 53 do Quadro 14, a Sentença 2 possui um advérbio (“inicialmente”) que evoca informação temporal. Dessa forma, os atributos “ocorrência de advérbio em S1” e “ocorrência de advérbio em S2” podem ter os seguintes valores: “n_temp”, “temp” ou “nsa”.

Para especificar o valor do atributo “ocorrência de ET em S1” e “ocorrência de ET em S2”, observou-se, com base na descrição das sentenças de um par, se havia a ocorrência de expressões de tempo, assinaladas por meio de sua categoria, ou não. O par 124 do Quadro 14, por exemplo, possui “ocorrência de ET em S1” com valor “data e hora” e o atributo “ocorrência de ET em S2” com valor “data”. Além dos valores que indicam as categorias das expressões que ocorrem nas sentenças, os referidos atributos podem ter o valor “nsa”, que indica a não ocorrência das informações linguísticas subjacentes e eles nas sentenças.

A especificação do atributo “sobreposição de subtópico” resulta da verificação da ocorrência de subtópicos idênticos ou diferentes entre as sentenças de um par, o que gera os valores binários possíveis “sim” ou “não”. Por exemplo, o par 6 do Quadro 14 é composto por sentenças que veiculam o mesmo subtópico, já que ambas, segundo a descrição das informações ilustradas no Quadro 14, expressam o subtópico “1”. Nesse caso, o valor especificado para o atributo em questão é “sim”.

Para especificar os atributos “ocorrência de marcador discursivo em S1” e “ocorrência de marcador discursivo em S2”, seguiu-se procedimento semelhante ao aplicado aos atributos “ocorrência de ET em S1” e “ocorrência de ET em S2”. No entanto, no caso dos marcadores, os valores possíveis podem ser os próprios marcadores constitutivos da sentença ou “nsa”. O par 6 do Quadro 14, por exemplo, possui o marcador discursivo “também” expresso na Sentença 1 e nenhum outro veiculado na Sentença 2, o que é expresso pelo valor “nsa”.

Os resultados da verificação ou cálculo dos atributos foram organizados em uma tabela no formato *xlsx* para que pudessem ser analisados de forma manual e também automática. A Tabela 11 ilustra essa organização. Nela, a denominação dos atributos foi abreviada: distância (Dist), sobreposição de nome (N), ocorrência de advérbio na sentença 1 (Adv_S1), ocorrência de advérbio na sentença 2 (Adv_S2), ocorrência de expressão temporal na sentença 1 (ET_S1), ocorrência de expressão temporal na sentença 2 (ET_S2), sobreposição de subtópico (SubT), ocorrência de marcador discursivo na sentença 1 (MD_S1) e ocorrência de marcador discursivo na sentença 2 (MD_S2).

Tabela 11: 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*	0.83	0	n_temp	nsa	nsa	data	Sim	também	nsa
57	10	*Historical background*	0.14	1	nsa	temp	nsa	nsa	Sim	nsa	nsa
124	3	*Elaboration*	0	3	nsa	nsa	data_hora	data	Sim	nsa	nsa

Fonte: Elaborado pelo autor.

Com base na Tabela 11, ressalta-se que:

(i) Os atributos “Dist” e “N” são numéricos, cujos valores reais podem variar entre 0 e 1;

(ii) Os atributos “Adv_S1” e “Adv_S2” são nominais ou categóricos, pois podem assumir um pequeno conjunto de valores possíveis, a saber: “temp” (advérbio temporal), “n_temp” (advérbio não-temporal) e “nsa”;

(iii) Os atributos “ET_S1 e “ET_S2” também são categóricos; no caso, os valores correspondem aos subtipos das ET, p.ex. “data”, “hora”, etc., além de “nsa”;

(iv) O atributo “SubT” também é categórico; no caso, apenas dois valores são possíveis, “sim” ou “não”, o que faz dele um atributo booleano ou binário;

(v) Os atributos “MD_S1 “MD_S2” também são categóricos; no caso, tais atributos podem ser especificados pelas próprias palavras da sentenças que funcionam como marcadores discursivos ou por “nsa”; a possibilidade de especificar os atributos por meio das próprias palavras que ocorrem na sentença faz deles atributos de valores lexicais.

Os resultados ilustrados na Tabela 11 foram analisados de forma manual e automática, com o objetivo de verificar a pertinência dos atributos para a caracterização do fenômeno da complementaridade, buscando reduzir o conjunto inicial de 9 atributos. A seleção manual e a automática foram feitas em função das relações CST e dos tipos de complementaridade.

4.2.1 Quanto à distinção das relações CST de complementaridade

A seleção manual dos atributos mais pertinentes para a distinção das relações partiu da verificação da frequência de ocorrência de cada um dos atributos linguísticos nos pares de sentença do *subcorpus* 1 em função das relações CST de complementaridade.

Diante do fato de que alguns atributos têm valores numéricos e outros têm valores categóricos, a frequência de ocorrência dos atributos foi feita de forma diferente em função do tipo de valor.

Para calcular a frequência de um atributo numérico *x*, (i) identificou-se a média simples dos valores (normalizados) de *x* obtidos para os 45 pares de cada relação CST de complementaridade e, na sequência, (ii) verificou-se, dentre os 45 pares, a quantidade cujo valor de *x* era igual ou superior à média.

Por exemplo, quanto ao atributo "distância", identificaram-se as seguintes médias: (i) 0,34 para os pares anotados com *Follow-up*; (ii) 0,53 para os pares anotados com *Historical background*, e (iii) 0,44 para os pares anotados com *Elaboration.* Em seguida, verificou-se que: (i) dos 45 pares de *Follow-up*, 19 possuem o atributo "distância" igual ou maior a média 0,34; (ii) dos 45 pares de *Historical background*, 23 deles possuem o valor do atributo "distância" igual ou maior que 0,53, e (iii) dos 45 pares de *Elaboration*, 20 possuem o valor desse atributo igual ou superior à média 0,44. Tal procedimento também foi adotado para a análise do atributo "sobreposição de nomes".

Quanto aos atributos categóricos, contabilizou-se a quantidade simples de pares em que o atributo ocorreu dentre os 45 de cada relação. Por exemplo, quanto ao atributo "advérbio em S1", verificou-se que: (i) dos 45 pares de sentenças anotados com a relação *Follow-up*, 5 deles possuíam advérbios temporais relacionados à complementaridade (isto é, 5/45); (ii) dos 45 pares anotados com *Historical background*, 13 deles possuíam advérbios temporais (isto é, 13/45), e (iii) dos 45 pares anotados com *Elaboration*, somente em 7 deles esse atributo foi observado (isto é, 7/45). Tal procedimento também foi adotado para a análise dos atributos "advérbio em S2", "expressão temporal em S1", "expressão temporal em S2", "sobreposição de subtópico", "ocorrência de marcador discursivo em S1" e "ocorrência de marcador discursivo em S2".

O resultado da verificação manual da frequência de cada atributo está descrito na Tabela 12.

Tabela 12: Seleção de atributos manual em função das relações CST de complementaridade.

Atributo	**Relação CST**		
	Follow-up	***Historical background***	***Elaboration***
Distância	19/45	23/45	20/45
Sobreposição de nome	22/45	24/45	27/45
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
Marcador discursivo em S1	10/45	6/45	8/45
Marcador discursivo em S2	2/45	2/45	3/45

Fonte: Elaborado pelo autor.

Com base na Tabela 12, observou-se que:

a) os atributos “distância” e “sobreposição de nome” parecem não diferenciar as relações de complementaridade entre si, pois a frequência desses atributos é elevada dentre os 45 pares de cada relação; a frequência de “sobreposição de nome”, aliás, é bastante similar entre os 45 pares referentes a cada relação CST (22/45, 24/45 e 27/45);

b) o atributo “advérbio em S1” destaca-se discretamente pela frequência mais elevada nos 45 pares anotados com *Historical background* em relação às outras duas relações;

c) o atributo “advérbio em S2” ocorre com baixíssima frequência nos pares anotados com *Follow-up* (1/45); entretanto, esse atributo também não é muito frequentes nos pares de *Historical background* (6/45) e *Elaboration* (5/45);

d) o atributo “expressão temporal em S1” é frequente nos pares anotados com *Historical background* (35/45) e nos pares anotados com *Follow-up* (18/45), indicando que esse atributo parece ser relevante para a discriminação da complementaridade temporal, já que, nos pares anotados com *Elaboration*, sua frequência é bem mais baixa (8/45);

e) o atributo “expressão temporal em S2” também não expressa a diferença entre as relações de complementaridade de forma significativa;

f) a frequência do atributo “sobreposição de subtópico” é elevada e similar nos pares de *Follow-up* (21/45) e *Elaboration* (22/45); ao contrário, esse atributo ocorre com frequência baixíssima nos pares de *Historical background*, o que pode indicar que a ausência desse atributo caracteriza a relação *Historical background*;

g) a frequência do atributo “marcador discursivo em S1” é discretamente mais elevada nos 45 pares anotados com a relação *Follow-up* (10/45);

h) o atributo “marcador discursivo em S2” tem frequência baixa nos pares anotados com as três relações CST de complementaridade, isto é, *Follow-up* (2/45), *Historical background* (2/45) e *Elaboration* e (3/45); dessa forma, esse atributo parece não as diferenciar as relações entre si.

A análise automática consistiu em verificar a pertinência dos atributos para diferenciar as relações CST de complementaridade por meio do algoritmo InfoGainAttributeEval, amplamente utilizado no PLN. Esse algoritmo, disponível no Weka, ranqueia os atributos em função do poder discriminativo de cada um deles e foi escolhido porque é capaz de analisar os atributos categóricos e numéricos, atribuindo-lhes pesos semelhantes. Para analisar os atributos, os dados completos, resultantes da verificação ou cálculo dos 9 atributos referentes aos pares de sentenças do *subcorpus* 1, foram submetidos ao InfoGainAttributeEval. Como resultado, o algoritmo organizou os 9 atributos de forma decrescente quanto à sua pertinência, como apresentado na Tabela 13.

Tabela 13: Seleção de atributos automática em função das relações CST de complementaridade.

Desempenho	**Atributo**
0.44	Expressão temporal em S1
0.07	Advérbio em S1
0.06	Marcador discursivo em S1
0.04	Marcador discursivo em S2
0.04	Sobreposição de subtópico
0.02	Advérbio em S2
0.01	Expressão temporal em S2
0	Sobreposição de nome
0	Distância

Os resultados da Tabela 13 confirmam algumas das observações feitas com base na análise manual da frequência de ocorrência dos atributos. Em especial, confirmam que o atributo “expressão temporal em S1” é o mais relevante e que os atributos “sobreposição de nome” e “distância” não são relevantes a tarefa questão.

Da seleção manual e automática, tem-se que o conjunto dos atributos mais relevantes é composto apenas por 7: “expressão temporal em S1”, “advérbio em S1”, “marcador discursivo em S1”, “marcador discursivo em S2”, “sobreposição de subtópico”, “advérbio em S2” e “expressão temporal em S2”.

4.2.2 Quanto à distinção dos tipos de complementaridade

A análise manual da pertinência dos atributos para distinguir os tipos de complementaridade também se baseou na frequência de ocorrência dos atributos. Tendo em vista o que já fora comentado sobre o desbalanceamento do *subcorpus* 1 no que diz respeito aos tipos de complementaridade, a frequência de ocorrência dos atributos foi transformada em valores percentuais, levando-se em conta esse desbalanceamento, para, assim, evitar discrepância entre os resultados. Por exemplo, os valores do atributo "distância" das relações *Follow-up* e *Historical background* (cf. Tabela 12) foram somados e transformados em valor percentual, resultando em 46% de pares de sentenças que possuem distâncias como indicativo de complementaridade temporal. Em outras palavras, em 46% dos pares de sentenças com complementaridade do tipo temporal o atributo "distância" é relevante para identificar a redundância entre elas. Demonstram-se os resultados dessa análise manual na Tabela 14.

Tabela 14: Seleção de atributos manual em função dos tipos de complementaridade.

Atributo	**Tipo de complementaridade**	
	Temporal	**Atemporal**
Distância	46%	44,4%
Sobreposição de nome	51,1%	60%
Advérbio em S1	20%	15,5%
Advérbio em S2	7,7%	11,11%
Expressão temporal em S1	58,8%	17,7%
Expressão temporal em S2	40%	37,7%
Sobreposição de subtópico	34,4%	48,8%
Marcador discursivo em S1	17,7%	17,7%
Marcador discursivo em S2	4,4%	6,6%

Fonte: Elaborado pelo autor.

Com base na Tabela 14, observa-se:

a) os atributos "distância" e "sobreposição de nome" ocorrem com frequência alta e similar nos pares com complementaridade temporal e atemporal; especificamente, o atributo "distância" ocorre em 46% dos pares do tipo temporal e em 44% dos pares do tipo atemporal; "sobreposição de nome", por sua vez, ocorre em 51% dos casos com complementaridade temporal e em 60% dos casos com complementaridade atemporal;

b) o atributo “expressão temporal em S1” ocorre com frequência alta nos pares com complementaridade temporal (ou seja, em 58% dos pares) e com baixa frequência nos pares anotados com complementaridade atemporal (17%);

c) o atributo “expressão temporal em S2” tem frequência relativamente alta e similar nos conjuntos de complementaridade; especificamente, o atributo “expressão temporal em S2” ocorre em 40% dos pares do tipo temporal e em 37,7% dos pares do tipo atemporal;

d) o atributo “sobreposição de subtópicos” ocorre em quase metade dos pares com complementaridade atemporal, ou seja, 48,8%; no conjunto com complementaridade temporal, a frequência é de 34,4%;

e) os atributos “advérbio em S1” e “advérbio em S2” possuem frequência relativamente baixa e similar em ambos os conjunto de complementaridade; especificamente, o atributo “advérbio em S1” ocorre em 20% dos pares do tipo temporal e em 15,5% dos pares do tipo atemporal; “advérbio em S2”, por sua vez, ocorre em 7,7% dos casos com complementaridade temporal e em 11,1% dos casos com complementaridade atemporal;

f) os atributos “marcador discursivo em S1” e “marcador discursivo em S2” também possuem frequência relativamente baixa e similar em ambos os conjunto de complementaridade; especificamente, o atributo “marcador discursivo em S1” ocorre com a mesma frequência em ambas as complementaridade, 17,7%; o atributo “marcador discursivo em S2”, por sua vez, ocorre em 4,4% dos casos com complementaridade temporal e em 6,6% dos casos com complementaridade atemporal.

Na seleção de atributos automática em função dos tipos de complementaridade, o *subcorpus* 1 também foi submetido algoritmo InfoGainAttributeEval, cujos resultados estão descritos na Tabela 15.

Tabela 15: Seleção de atributos automática em função dos tipos de complementaridade.

Desempenho	Atributo
0.17	Expressão temporal em S1
0.02	Marcador discursivo em S1
0.01	Marcador discursivo em S2
0.01	Sobreposição de subtópico
0.01	Expressão temporal em S2
0.002	Advérbio em S1
0.002	Advérbio em S2
0	Sobreposição de nomes
0	Distância

Fonte: Elaborado pelo autor.

Nessa Tabela, tem-se o ranqueamento produzido pelo InfoGainAttributeEval para os 9 atributos inicialmente delimitados. Esse ranque revela a pertinência estatística de cada atributo individual para identificar os diferentes tipos de complementaridade. No caso, observa-se que "expressão temporal em S1" (0.17) é a atributo mais relevante para distinguir os tipos de complementaridade. O desempenho de "marcador discursivo em S1" (0.02), "marcador discursivo em S2" (0.01), "sobreposição de subtópico" (0.01) e "expressão temporal em S2" (0.01) foi similar e abaixo do obtido por "expressão temporal em S1". Os atributos "advérbio em S1" e "advérbio em S2" tiveram resultados bem inferiores (0.002) e "sobreposição de nomes" e "distância", por sua, não são relevantes para a distinção segundo o aprendizado estatístico, já que obtiveram "0" de relevância. Diante desse ranqueamento, identificou-se também um subconjunto de 7 atributos efetivamente relevantes, excluindo-se, portanto, "sobreposição de nomes" e "distância".

No próximo Capítulo, apresentam-se o teste e a avaliação dos atributos por meio de algoritmos supervisionados de AM com vistas a determinar a relevância dos atributos para distinguir as relações CST e dos tipos de complementaridade. Para tanto, a unificação dos *subcorpora* 1 e 2 foi submetida ao AM. Ademais, salienta-se que somente o conjunto de 7 atributos resultante do processo de seleção de atributos foi testado nas referidas tarefas. A escolha pelo AM pautou-se basicamente no fato de que o aprendizado estatístico pode correlacionar os atributos, gerando as combinações que mais adequadamente capturam o fenômeno em questão.

Capítulo 5

TESTE E AVALIAÇÃO PARA IDENTIFICAÇÃO DAS RELAÇÕES E TIPOS DE COMPLEMENTARIDADE

5.1 Especificações para o aprendizado de máquina

Para investigar os atributos, selecionaram-se os algoritmos supervisionados utilizados por Maziero (2012) e Souza *et al.* (2012), a saber: (i) PART (WITTEN, FRANK, 1998), (ii) J48 (QUILAN, 1993), (iii) OneR (HOLTE, 1993). Como mencionado, os algoritmos do paradigma supervisionado adquirem conhecimento implícito de exemplos previamente classificados, gerando classificadores (p.ex.: conjunto de regras) que relacionam os atributos (e seus valores) às classes. Em outras palavras, esses algoritmos partem de *corpora* anotados com classes (no caso, informação linguística a ser aprendida) e atributos (isto é, características das instâncias que potencialmente são relevantes para o aprendizado estatístico das classes) e capturam padrões que evidenciam o(s) atributo(s) ou a combinação de atributos que caracterizam cada uma das classes. Tais padrões subsidiam os chamados classificadores.

Assim, o algoritmo PART analisa um conjunto (ou *corpus*) de instâncias (no caso, os pares de sentenças), às quais estão associadas de forma explícita: (i) as classes que se quer aprender estatisticamente, no caso, as relações CST e os tipos de complementaridade, e (ii) os atributos/valores cuja pertinência para determinar as classes se quer testar. Tal conjunto de dados é ilustrado pela Tabela 11. Como resultado da análise, o PART gera classificadores que se constituem de regras no formato lógico, que comumente combinam atributos para capturar mais adequadamente as classes.

O algoritmo J48 analisa o conjunto de dados de maneira bastante semelhante ao PART, mas o resultado de sua análise é disponibilizado no formato de árvores de decisão. A abordagem utilizada durante a construção da árvore é *top-down*, em que o atributo mais significativo (logo, o mais genérico) é tido como o nó inicial da árvore. Os nós seguintes da árvore são os menos significativos em comparação aos anteriores.

Por fim, o algoritmo OneR baseia-se em regras de decisão e gera uma única regra como resultado de sua análise estatística. O algoritmo pode realizar predições estatísticas valendo-se de atributos categóricos e/ou numéricos, atribuindo-lhes pesos de igual valor.

Os 3 algoritmos foram aplicados para testar os atributos em um *corpus* que, como mencionado, é a unificação dos *subcorpora* 1 e 2. Assim, os algoritmos foram testados em um conjunto de sentenças constituído por (i) pares utilizados para a descrição manual do fenômeno e a seleção de atributos manual e automática (*subcorpus* 1) e (ii) um novo grupo de pares (*subcorpus* 2), até então não utilizado para o estudo do fenômeno. Para tanto, realizou-se o pré-processamento do *subcorpus* 2, segundo as diretrizes aplicadas ao *subcorpus* 1, ou seja: (i) descrição semiautomática das informações linguísticas de cada sentença dos pares subjacentes aos atributos e (ii) verificação/cálculo manual dos atributos para cada par. Ao final, os algoritmos foram aplicados em um conjunto de pares de sentenças distribuídas por relações da seguinte maneira: 114 pares de sentenças com *Elaboration*, 106 pares com a relação *Follow-up* e 62 de *Historical background.*

Especificamente, ressalta-se que apenas o conjunto de 7 atributos, resultante do processo de "seleção de atributos", foram testados pelos algoritmos. Nesse caso, diz-se que os referidos teste e avaliação foram feitos com seleção de conteúdo.

Ademais, salienta-se que o teste dos 7 atributos baseou-se na amostragem *10-fold cross-validation* (isto é, "validação cruzada de 10 pastas"), amplamente aplicada quando o aprendizado é feito com uma quantidade pequena de dados, como é o caso do *subcorpus*1+2 sob análise. Essa técnica caracteriza-se por particionar o conjunto de dados (*subcorpus*1+2) em 10 subconjuntos mutualmente exclusivos, sendo que a cada execução do algoritmo de AM são utilizados 9 subconjuntos para classificar e somente 1para validar o classificador (ou algoritmo). Esse particionamento simula, então, uma situação real, em que os dados submetidos ao treinamento e teste são disjuntos.

Por fim, o aprendizado com seleção de conteúdo, gerado por cada um dos 3 algoritmos para as tarefas de distinção das relações CST e dos tipos de complementaridade, foi avaliado em função das medidas de avaliação tradicionais da SA, ou seja, precisão, cobertura e medida-f.

5.2 Correlação entre os atributos e as relações CST de complementaridade

O primeiro algoritmo utilizado para testar os 7 atributos foi o PART, que aprendeu as 21 regras lógicas (formato "*se*, *então*") do Quadro 15 para detecção das relações CST de complementaridade. A esse conjunto de regras é dado o nome de classificador.

Quadro 15: Regras do PART para distinção das relações CST.

1. Se ET_S1 = hora, **então** *Follow-up*
2. Senão ET_S1 = data **E** ADV_S1 = nsa **E** ET_S2 = nsa, **então** *Historical background*
3. Senão MD_S2 = "também", **então** *Follow-up*
4. Senão ET_S1 = absoluto, **então** *Historical background*
5. Senão ET_S1 = nsa **E** MD_S1 = "também" **E** ADV_S1 = n_temp, **então** *Follow-up*
6. Senão ET_S1 = nsa **E** ET_S2 = enunciado **E** SUBT = sim, **então** *Elaboration*
7. Senão ET_S1 = nsa **E** MD_S2 = nsa **E** ET_S2 = enunciado **E** ADV_S2 = nsa, **então** *Elaboration*
8. Senão ET_S1 = nsa **E** MD_S2 = nsa **E** ET_S2 = nsa, **então** *Elaboration*
9. Senão MD_S2 = nsa **E** MD_S1 = "também" **E** SUBT = não, **então** *Historical background*
10. Senão MD_S2 = nsa **E** ET_S1 = enunciado **E** ADV_S1 = nsa, **então** *Follow-up*
11. Senão MD_S2 = nsa **E** MD_S1 = nsa **E** ET_S2 = enunciado, **então** *Follow-up*
12. Senão MD_S2 = nsa **E** ET_S1 = nsa **E** SUBT = sim, **então** *Elaboration*
13. Senão MD_S2 = nsa **E** ET_S2 = hora, **então** *Historical background*
14. Senão MD_S2 = nsa **E** ET_S2 = data **E** ADV_S2 = n_temp **E** ADV_S1 = nsa, **então** *Follow-up*
15. Senão MD_S2 = nsa **E** ET_S2 = data **E** ADV_S2 = nsa **E** SUBT = não, **então** *Historical background*
16. Senão MD_S2 = nsa **E** ADV_S1 = n_temp, **então** *Elaboration*
17. Senão MD_S2 = nsa **E** ADV_S1 = temp **E** ET_S1 = nsa, **então** *Elaboration*
18. Senão MD_S2 = nsa **E** ADV_S1 = nsa **E** SUBT = sim, **então** *Follow-up*
19. Senão MD_S2 = nsa **E** ET_S1 = data **E** ADV_S1 = temp, **então** *Historical background*
20. Senão MD_S2 = nsa, **então** *Elaboration*
21. Caso contrário, *Follow-up*

Fonte: Elaborado pelo autor.

A Regra 1, por exemplo, estabelece que, dado um par de sentenças S1 e S2, caso ocorra uma expressão temporal do tipo hora na S1, a relação CST que se estabelece entre S1 e S2 é *Follow-up*. Como a aplicação das regras é sequencial, a Regra 2 é empregada somente quando a Regra 1 não é satisfeita, ou seja, quando o atributo ET_S1 tiver valor diferente de "hora", e quando os atributos e valores que constituem a Regra são satisfeitos (isto é, ET_S1=data, ADV_S1=nsa e ET_S2=nsa). Assim, a Regra só é aplicada quando a (i) S1 não possuir uma ET do tipo "hora" (negação da Regra 1), mas sim de "data", e não possuir um advérbio de tempo, e quando a (ii) S2 não possuir uma ET. Diante disso, a relação entre S1 e S2 é de

Historical background. Quando todas as 20 regras não são satisfeitas, o PART identificou que a relação é *Follow-up*, o que está codificado na Regra 21.

As 21 regras geradas pelo PART obtiveram o desempenho apresentado na Tabela 16, codificado pelas medidas clássicas de precisão, cobertura e medida-f.

Tabela 16: Avaliação das regras do PART para identificação das relações CST.

Relação CST	**Precisão**	**Cobertura**	**Medida-F**
Historical background	0.75	0.75	0.75
Follow-up	0.71	0.6	0.65
Elaboration	0.59	0.68	0.63

Fonte: Elaborado pelo autor.

Observa-se que a detecção correta dos pares de *Historical background* obteve a medida-f como a mais alta (75%) frente aos pares das demais relações, cuja identificação obteve a mesma medida bastante similar. No caso, *Follow-up* obteve medida-f de 65% e *Elaboration* de 63%.

No Quadro 16, tem-se a matriz de confusão resultante da aplicação das regras, a qual oferece uma medida efetiva do modelo de classificação ao mostrar o número de classificações corretas *versus* as classificações preditas para cada classe, sobre um conjunto de exemplos.

Quadro 16: Matriz de confusão das regras do PART para identificação das relações CST.

Classe / **Teste**	*Follow-up* (106)	*Historical background* (62)	*Elaboration* (114)
Follow-up	64	3	39
Historical background	1	47	14
Elaboration	24	12	78

Fonte: Elaborado pelo autor.

No Quadro 16, observa-se que o classificador confunde pouco as relações de complementaridade temporal entre si. Dos 106 pares com relação *Follow-up* do *subcorpus* 1+2, 64 foram identificadas corretamente e 3 pares foram identificados como *Historical background.* Além disso, dos 62 pares de *Historical background,* 47 foram corretamente identificados pelo classificador e apenas 1 par foi erroneamente identificado como *Follow-up*. Quanto à *Elaboration*, as regras a confundem mais vezes com *Follow-up*, já que, dos 114 pares de *Elaboration,* 24 foram classificados erroneamente como *Follow-up* e somente 12 como *Historical background*.

Da matriz, pode-se dizer ainda que o classificador em questão diferencia de maneira razoável as relações, acertando, para as 3 relações, um pouco mais de 50% dos pares: (i) 64 pares corretos de *Follow-up* equivalem a 60% de 106, (ii) 47 pares corretos de *Historical background* equivalem a 75% de 62, e (iii) 78 pares corretos de *Elaboration* equivalem a 68% de 114.

O algoritmo OneR, por sua vez, gerou a regra apresentada do Quadro 17, a qual se baseia exclusivamente no atributo "expressão temporal em S1".

Quadro 17: Regra do OneR para identificação das relações CST.

ET_S1:
- nsa → *Elaboration*
- data → *Historical background*
- hora → *Follow-up*
- duração → *Elaboration*
- enunciado → *Follow-up*
- absoluto → *Historical background*

Fonte: Elaborado pelo autor.

Na Tabela 17, registra-se o desempenho do classificador do OneR.

Tabela 17: Avaliação da regra do OneR para identificação das relações CST.

Relação CST	**Precisão**	**Cobertura**	**Medida-F**
Historical background	0.76	0.72	0.74
Follow-up	0.73	0.43	0.54
Elaboration	0.58	0.82	0.68

Fonte: Elaborado pelo autor.

Na Tabela 17, observa-se que o classificador do OneR, assim como o do algoritmo PART, obteve a medida-f como a mais alta quanto à identificação da relação *Historical background* (74%). Essa medida, aliás, é bastante similar à obtida pelo PART para a mesma relação, que foi de 75%. Na sequência, o segundo melhor resultado com a mesma medida diz respeito à detecção da relação *Elaboration* (68%). *Follow-up*, por sua vez, obteve o resultado mais baixo, 58% somente. A título de comparação, ressalta-se que PART, ao contrário do OneR,

obteve medida-f bastante similar para a identificação das relações *Follow-up* e *Elaboration*, 65% e 63%, respectivamente.

No Quadro 18, apresenta-se a matriz de confusão gerada pela regra aprendida pelo algoritmo OneR.

Quadro 18: Matriz de confusão do OneR para identificação das relações CST.

Classe / **Teste**	*Follow-up* (106)	*Historical background* (62)	*Elaboration* (114)
Follow-up	46	6	54
Historical background	5	45	12
Elaboration	12	8	94

Fonte: Elaborado pelo autor.

Com base no Quadro 18, vê-se que o OneR identifica a relação *Elabaration* com precisão mais alta que o PART, já que classificou corretamente 94 pares frente aos 78 do PART. Quanto à *Follow-up*, observa-se que essa relação é fortemente confundida com *Elaboration*, já que, dos 106 pares do *subcorpus* 1+2, 46 foram detectados corretamente e 54 foram classificados erroneamente como *Elaboration.* Nesse caso, o PART faz menos confusão entre as duas, acertando 64 dos 106 de *Follow-up*. Sobre *Historical background*, o classificador do OneR obteve desempenho bastante semelhante ao do PART, pois acertou 45 dos 62 pares, quando o PART acertou 47.

Por fim, os mesmos 7 atributos foram testados pelo algoritmo J48, que produz regras no formato de árvore de decisão. Tais regras resultam da correlação entre o atributo mais recorrente entre as instâncias e os demais.

No Quadro 19, apresentam-se as regras do J48. Nelas, observa-se, por exemplo, que a parte inicial da árvore estabelece que, se um par de sentenças possui os atributos “ET_S1”, “MD_S2” e “MD_S1” com valor “nsa”, a relação será *Elaboration*. Diante da estrutura da árvore, vê-se que o atributo mais recorrente é “expressão temporal em S1” (ET_S1), pois este ocupa o topo, sendo comum a todos os ramos da árvore.

Ainda quanto às regras do J48, destaca-se o fato de o atributo “sobreposição de subtópico” não compor a árvore de decisão, o qual havia sido identificado como o quarto mais relevante na seleção de atributos. Em contrapartida, o conjunto de regras desse algoritmo se baseia em “expressões temporais em S2” e “advérbio em S1” que, segundo o ranqueamento dos atributos, possuem relevância igual ou inferior a “sobreposição de subtópico”.

Quadro 19: Regras do J48 para identificação das relações CST de complementaridade.

```
ET_S1 = nsa
|   MD_S2 = nsa
|   |   MD_S1 = nsa: Elaboration
|   |   MD_S1 = “também”
|   |   |   ADV_S1 = nsa: Elaboration
|   |   |   ADV_S1 = n_temp: Follow-up
|   |   |   ADV_S1 = temp: Follow-up
|   |   |   ADV_S1 = temp : Follow-up
|   |   MD_S1 = “ainda”: Elaboration
|   |   MD_S1 = “além_de”: Elaboration
|   |   MD_S1 = “onde”: Follow-up
|   |   MD_S1 = “após”: Elaboration
|   MD_S2 = “por_exemplo”: Follow-up
|   MD_S2 = “também”: Follow-up
|   MD_S2 = “ainda”: Elaboration
|   MD_S2 = “após”: Elaboration
|   MD_S2 = “além_de”: Follow-up
|   MD_S2 = “além_de”: Follow-up
|   MD_S2 = “onde”: Follow-up
ET_S1 = data: Historical background
ET_S1 = hora: Follow-up
ET_S1 = duração
|   ADV_S1 = nsa: Elaboration
|   ADV_S1 = n_temp: Historical background
|   ADV_S1 = temp: Elaboration
|   ADV_S1 = temp : Elaboration
ET_S1 = enunciado
|   MD_S1 = nsa: Follow-up
|   MD_S1 = “também”
|   |   ET_S2 = nsa: Historical background
|   |   ET_S2 = data: Historical background
|   |   ET_S2 = hora: Historical background
|   |   ET_S2 = enunciado: Follow-up
|   |   ET_S2 = intervalo: Historical background
|   |   ET_S2 = absoluto: Historical background
|   |   ET_S2 = duração: Historical background
|   MD_S1 = “ainda”: Follow-up
|   MD_S1 = “além_de”: Follow-up
|   MD_S1 = “onde”: Follow-up
|   MD_S1 = “após”: Follow-up
ET_S1 = absoluto: Historical background
```

Fonte: Elaborado pelo autor.

Na Tabela 18, apresenta-se o desempenho do conjunto de regras (ou classificador) aprendido pelo algoritmo J48 para identificar as relações de complementaridade.

Tabela 18: Avaliação das regras do J48 para identificação das relações CST.

Relação CST	Precisão	Cobertura	Medida-F
Historical background	0.78	0.75	0.75
Follow-up	0.7	0.5	0.58
Elaboration	0.6	0.76	0.67

Fonte: Elaborado pelo autor.

De acordo com a Tabela 18, é possível observar que o desempenho do J48 é bastante similar ao obtido pelo classificador do algoritmo OneR. Isso se justifica porque ambos obtiveram 75% de medida-f ao identificar a relação *Historical background.* Além disso, a medida-f para a detecção de *Elaboration* é basicamente a mesma, já que o J48 obteve 67% e o OneR obteve 68%. Por fim, ao comparar o desempenho dos algoritmos, observa-se que as regras do J48 obtiveram uma medida-f sutilmente mais alta que o OneR para a identificação das relações de *Follow-up*, 58% e 54%, respectivamente.

No Quadro 20, tem-se a matriz de confusão gera pela árvore de decisão do J48.

Quadro 20: Matriz de confusão do J48 para identificação das relações CST.

Classe / **Teste**	*Follow-up* (106)	*Historical background* (62)	*Elaboration* (114)
Follow-up	53	7	46
Historical background	3	47	12
Elaboration	19	8	87

Fonte: Elaborado pelo autor.

De acordo com a matriz do Quadro 20, o J48 identifica corretamente os mesmos 47 pares de *Historical background* (de 62) que o PART, classificando erroneamente, no entanto, mais pares como *Follow-up* que o PART (3 *versus* 1). Para a identificação de *Follow-up,* o desempenho do J48 é o segundo melhor, já que o PART identificou 64 pares dos 106, o J48 detectou 53 e o OneR, por sua vez, 46. A maior confusão ao classificar *Follow-up,* ocorre com *Elabaration,* como ocorre com os demais algoritmos. Ainda sobre a relação *Elaboration*, o algoritmo classifica corretamente 87 pares de sentenças, mas equivoca-se com a relação *Follow-up* em 19 pares. Por fim, sabe-se que a quantidade de classificação incorreta para a relação *Historical background* é menor,

em que confunde-se apenas 8 pares de sentenças ao tentar classificar a relação *Elaboration*.

A seguir, relata-se a investigação da pertinência dos atributos para a detecção dos tipos de complementaridade.

5.3 Análise da correlação entre os atributos e os tipos de complementaridade

O potencial do conjunto de 7 atributos, como mencionado, também foi testado para a detecção dos tipos de complementaridade, seguindo-se as mesmas diretrizes aplicadas à investigação dos atributos para a distinção das relações CST.

No Quadro 21, exibe-se o conjunto de regras lógicas aprendidas pelo PART, as quais constituem o classificador desse algoritmo.

Quadro 21: Regras do PART para identificação dos tipos de complementaridade.

1. Se ET_S1 = data, **então** Complementaridade Temporal

2. Se ET_S1 = enunciado, **então** Complementaridade Temporal

3. Se ET_S1 = hora, **então** Complementaridade Temporal

4. Se MD_S2 = nsa **E** ET_S1 = nsa **E** MD_S1 = nsa **E** ET_S2 = enunciado **E** SUBT = sim, **então** Complementaridade Atemporal

5. Se MD_S2 = também, **então** Complementaridade Temporal

6. Se ET_S1 = absoluto, **então** Complementaridade Temporal

7. Se MD_S2 = nsa **E** MD_S1 = também **E** ADV_S1 = n_temp, **então** Complementaridade Temporal

8. Se MD_S2 = nsa **E** MD_S1 = nsa **E** ADV_S2 = nsa, **então** Complementaridade Atemporal

9. Se MD_S2 = nsa **E** MD_S1 = também, **então** Complementaridade Atemporal

10. Se MD_S2 = nsa **E** MD_S1 = ainda, **então** Complementaridade Atemporal

11. Se MD_S2 = nsa **E** MD_S1 = nsa **E** ET_S2 = nsa **E** ADV_S2 = temp, **então** Complementaridade Atemporal

12. Se MD_S2 = nsa **E** MD_S1 = nsa **E** ET_S2 = nsa **E** ADV_S2 = n_temp **E** SUBT = não, **então** Complementaridade Temporal

13. Se MD_S2 = nsa **E** SUBT = sim, **então** Complementaridade Atemporal

14. Se MD_S2 = nsa **E** ADV_S2 = n_temp, **então** Complementaridade Temporal

15. Se ET_S2 = nsa **E** ADV_S1 = nsa, **então** Complementaridade Atemporal

16. Se ADV_S1 = n_temp, **então** Complementaridade Temporal

17. Se ADV_S1 = nsa, **então** Complementaridade Temporal

18. caso contrário, Complementaridade Atemporal

Fonte: Elaborado pelo autor.

Ressalta-se que o PART gerou um conjunto de 18 regras lógicas para a tarefa em questão. Por exemplo, a Regra 1 estabelece que, dado um par de sentenças S1 e S2, se o atributo "ET_S1" obtiver valor "data", o tipo de complementaridade se estabelece entre S1 e S2 é temporal.

Na Tabela 19, têm-se os resultados da avaliação do classificador do PART.

Tabela 19: Avaliação das regras do PART para identificação dos tipos de complementaridade.

Tipo de complementaridade	Precisão	Cobertura	Medida-F
Temporal	0.76	0.73	0.74
Atemporal	0.62	0.66	0.64

Fonte: Elaborado pelo autor.

As medidas de avaliação apontam que a identificação da complementaridade temporal obteve medida-f mais alta que a complementaridade atemporal, 74% e 64%, respectivamente. Os valores obtidos para precisão e cobertura também são bastante similares entre aos tipos.

O desempenho geral do classificador pode ser resultado de dois fatores em especial: (i) maior número de instâncias de complementaridade temporal em detrimento da atemporal (168 e 114, respectivamente), o que permite um aprendizado mais adequado dos padrões estatísticos e, (ii) a característica mais linguisticamente marcada da complementaridade temporal frente atemporal, que é codificada por meio de um número maior de atributos, facilitando o aprendizado estatístico.

No Quadro 22, tem-se a matriz de confusão das regras do PART.

Quadro 22: Matriz de confusão do PART para identificação dos tipos de complementaridade.

Classe / **Teste**	Temporal (168)	Atemporal (114)
Temporal	123	45
Atemporal	38	76

Fonte: Elaborado pelo autor.

Para identificar os tipos de complementaridade, o OneR também produz uma regra baseada no atributo "expressão temporal em S1" (ET_S1), como evidenciado no Quadro 23.

Quadro 23: Regra do OneR para identificação dos tipos de complementaridade.

```
ET_S1:
        nsa          → Atemporal
        data         → Temporal
        hora         → Temporal
        duração      → Temporal
        enunciado → Temporal
        absoluto     → Temporal
```

Fonte: Elaborado pelo autor.

Na Tabela 20, apresentam-se os resultados obtidos na avaliação do classificador do OneR, composto pela regra do Quadro 23, para a identificação dos tipos de complementaridade.

Tabela 20: Avaliação do OneR para identificação dos tipos de complementaridade.

Tipo de complementaridade	**Precisão**	**Cobertura**	**Medida-F**
Temporal	0.82	0.6	0.69
Atemporal	0.58	0.8	0.67

Fonte: Elaborado pelo autor.

De acordo com a Tabela 20, o classificador do OneR obteve medida-f similar para os diferentes tipos de complementaridade, a saber: 69% para complementaridade temporal e 67% para complementaridade temporal. Quanto à medida-f, o desempenho do OneR é distinto do PART, já que a detecção do tipo temporal pelo PART obteve medida-f relativamente mais alta que a obtida para o tipo atemporal (74% e 64%, respectivamente). Com relação à precisão, salienta-se que o classificador do OneR, identifica o tipo temporal com precisão bem mais alta que o tipo atemporal (82% e 58%, respectivamente). No entanto, ocorre o inverso com a cobertura: 60% para temporal e 80% para atemporal.

Além de a medida-f demonstrar que a quantidade e a qualidade da identificação dos tipos de complementaridade estão equilibradas, a matriz de confusão, no Quadro 23, aponta que o classificador em questão comete menos equívocos para classificação.

Quadro 24: Matriz de confusão do OneR para identificação dos tipos de complementaridade.

Classe / **Teste**	Temporal (168)	Atemporal (114)
Temporal	102	66
Atemporal	22	92

Fonte: Elaborado pelo autor.

Segundo o Quadro 23, o OneR comente menos equívocos para a classificação dos tipos de complementaridade. Do conjunto de 168 pares de sentenças do tipo temporal, o classificador detecta 66 pares como complementaridade atemporal. Já dos 114 pares de sentenças do tipo atemporal que compõem o *subcorpus* 1+2, somente 22 pares são apontados como sendo de complementaridade temporal.

Por fim, apresenta-se, no Quadro 25, o conjunto de regras, no formato de árvore de decisão, aprendidas pelo J48 para detectar os diferentes tipos de complementaridade.

Quadro 25: Regras do J48 para identificação dos tipos de complementaridade.

```
ET_S1 = nsa
|   MD_S2 = nsa
|   |   MD_S1 = nsa: Atemporal
|   |   MD_S1 = "também"
|   |   |   ADV_S1 = nsa: Atemporal
|   |   |   ADV_S1 = n_temp: Temporal
|   |   |   ADV_S1 = temp: Temporal
|   |   |   ADV_S1 = temp: Temporal
|   |   MD_S1 = "ainda": Atemporal
|   |   MD_S1 = "além_de": Atemporal
|   |   MD_S1 = "onde": Temporal
|   |   MD_S1 = "após": Atemporal
|   MD_S2 = "por_exemplo": Temporal
|   MD_S2 = "também": Temporal
|   MD_S2 = "ainda": Atemporal
|   MD_S2 = "após": Atemporal
|   MD_S2 = "além_de": Temporal
|   MD_S2 = "onde": Atemporal
ET_S1 = data: Temporal
ET_S1 = hora: Temporal
ET_S1 = duração
|   ADV_S1 = nsa: Atemporal
|   ADV_S1 = n_temp: Temporal
|   ADV_S1 = temp: Atemporal
|   ADV_S1 = temp: Temporal
ET_S1 = enunciado: Temporal
ET_S1 = absoluto: Temporal
```

Fonte: Elaborado pelo autor.

Na árvore de decisão do Quadro 25, o atributo "expressão temporal em S1" (ET_S1), assim como na árvore gerada para as relações CST, é o mais recorrente e apontado

como o atributo distintivo entre os tipos de complementaridade, sendo, por isso, o nó raiz da árvore.

As regras obtidas pelo algoritmo não utilizam o atributo "sobreposição de subtópico", mas utilizam o atributo "advérbio em S1" que, de acordo com o algoritmo InfoGainAttributeEval, é o segundo atributo mais relevante em identificar (e distinguir) o tipo de complementaridade.

Quanto à identificação dos tipos de complementaridade, a árvore de decisão do Quadro 25 obteve o desempenho descrito na Tabela 21.

Tabela 21: Avaliação das regras do J48 para identificação dos tipos de complementaridade.

Tipos de complementaridade	Precisão	Cobertura	Medida-F
Temporal	0.8	0.68	0.74
Atemporal	0.61	0.75	0.68

Fonte: Elaborado pelo autor.

A diferença entre a medida-f obtida para a complementaridade temporal e para a atemporal (74% e 68%, respectivamente) não é tão grande quanto a do PART (74% e 64%, respectivamente), nem tão pequena quanto a do OneR (69% e 67%, respectivamente). Com relação à precisão e cobertura referentes à identificação do tipo temporal, ressalta-se que os valores do J48 (80% e 68%, respectivamente) são mais próximos dos obtidos pelo OneR (82% e 60%, respectivamente).

No Quadro 27, apresenta-se a matriz de confusão do J48.

Quadro 26: Matriz de confusão do J48 para identificação dos tipos de complementaridade.

Classe / **Teste**	Temporal (168)	Atemporal (114)
Temporal	115	53
Atemporal	28	86

Fonte: Elaborado pelo autor.

Segundo o Quadro 27, o algoritmo J48 identifica melhor as instâncias com complementaridade temporal. Entretanto, ao classificar esse tipo de complemento, o algoritmo comete equívocos, classificando 53 instâncias com complementaridade atemporal.

Por fim, é possível fazer um comparativo geral entre os classificadores gerados pelos 3 algoritmos supervisionados de AM em função de cada tarefa desempenhada:

a) o algoritmo PART, ao classificar as relações CST de complementaridade, confunde as relações *Follow-up* e *Elaboration*, mas é capaz de distinguir a relação *Historical background* das demais (cf. Quadro 16, pág. 85). Além disso, as 3 medidas de avaliação são maiores para a relação *Historical background*, e menores para a relação *Elaboration* (cf. Tabela 16, pág. 85). Por fim, o conjunto de regras é baseado em todos os atributos propostos após a seleção daqueles que eram os mais discriminativos (cf. Quadro 15, pág. 84), em especial o atributo ET_S1;
b) o algoritmo J48 também comete equívocos na classificação dos pares de sentenças das relações *Follow-up* e *Elaboration*, mas também consegue distinguir com maior precisão a relação *Historical background* (cf. Quadro 20, pág. 89), cujas medidas de avaliação são maiores em relação às demais relações (cf. Tabela 18, pág. 89). Sobre o conjunto de regras gerado, o algoritmo não utiliza os atributos SUBT e ADV_S2, apesar de classificar corretamente muitos pares de sentenças. Com relação ao algoritmo PART, J48 também aponta que o atributo ET_S1 é o mais discriminativo dentre os demais, considerando, inclusive, como o nó principal da árvore de decisão (cf. Quadro 19, pág. 88);
c) o algoritmo OneR, apesar de uma aparente simplicidade, é bastante poderoso. Ele também confunde as relações *Follow-up* e *Elaboration*, classificando com mais facilidade a relação *Historical background* (cf. Quadro 18, pág. 87). Acerca das medidas de avaliação, a medida "cobertura" destaca-se para a relação *Elaboration*, e as demais medidas para a relação *Historical background* (cf. Tabela 17, pág. 86). Por fim, o conjunto de regras baseia-se somente no atributo mais proeminente às relações (ET_S1) podendo ser considerado o mais discriminativo (cf. Quadro 17, pág. 86). Ainda sobre o conjunto de regras, é importante destacar que aqui está a simplicidade do algoritmo (utilizar só um atributo), mas também está o seu potencial, já que a quantidade de instâncias (pares de sentenças) corretamente classificadas aproxima-se dos outros algoritmos utilizados.

No próximo Capítulo, apresentam-se as considerações finais sobre esta pesquisa, além dos trabalhos futuros e contribuições.

Capítulo 6

CONSIDERAÇÕES FINAIS

6.1 Considerações gerais da pesquisa

Neste trabalho, realizou-se a primeira investigação sobre o fenômeno da complementaridade entre sentenças advindas de textos distintos que abordam mesmo assunto. Essa investigação objetivou não só descrever e compreender o fenômeno, por meio do levantamento de suas características linguísticas, como também traduzir tais características em atributos mais adequados para subsidiar a detecção automática das relações do modelo CST que codificam esse fenômeno como os tipos de complementaridade subjacentes a essas relações. Até então, as referidas relações CST têm sido automaticamente identificadas com certa precisão somente com base em atributos que capturam a redundância.

Diante da ausência de trabalhos específicos de caracterização do fenômeno em questão, foi necessário estudar sua manifestação e comportamento linguístico. Assim, o estudo descritivo com base em *corpus* foi de suma importância. Salienta-se que se o *corpus* não atende a parâmetros bem delimitados e não possui boa representatividade do fenômeno é impossível que a descrição do fenômeno seja a mais próxima de sua manifestação em uma produção textual comum e cotidiana. Dessa forma, o *corpus* CSTNews destaca-se por representar de maneira significativa a complementaridade, além de contar com camadas de anotações que foram úteis para a identificação de atributos relevantes para caracterização do fenômeno.

Se a descrição partir de um conjunto de textos que manifesta o fenômeno em seu ambiente natural de ocorrência, logo os atributos a serem propostos devem refletir sua realidade linguística. Assim, os atributos que identificam a redundância (sobreposição de nomes, distância e sobreposição de subtópico) podem ser considerados quanto essa realidade linguística presente nos pares de sentenças anotados com as relações de complementaridade. Além disso, por partir da tipologia que organiza as relações CST, proposta ao PB, consideraram-se atributos que fossem capazes detectar aspectos temporais (advérbio em S1,

advérbio em S2, expressões temporais em S1 e expressões temporais em S2) e atemporais (marcadores discursivos em S1 e marcadores discursivos em S2).

Mas, se por um lado a descrição deva partir da realidade/materialidade linguística, por outro, quando submetidos a ambientes de AM, os atributos linguísticos podem não ter resultados tão significativos na tarefa desenvolvida. Por conta disso, foi importante evidenciar quais atributos caracterizavam melhor a complementaridade. Dessa forma, selecionaram-se aqueles que possuíam desempenho mais significativo (como é o caso do atributo "expressão temporal em S1"), excluindo-se os atributos "sobreposição de nomes" e "distância" da análise.

Ainda sobre as técnicas de AM, elas desempenharam papel fundamental nessa pesquisa ao corroborar os apontamentos já realizados anteriormente de forma manual. Sabia-se que alguns atributos poderiam não ser discriminativos em função das relações CST de complementaridade e seus tipos. Tal apontamento foi confirmado ao utilizar a seleção de atributos. Além disso, a análise manual é limitada quanto a combinatória de atributos. Analisava-se um atributo por vez, sem saber, no entanto, que quando dado atributo ocorria em consonância com outro dado atributo poderia se estabelecer um padrão característico a uma relação. Isso também foi possível com o uso de algoritmos de AM.

As regras geradas pelo algoritmo ainda podem alcançar desempenho melhor. No entanto, é preciso voltar aos casos em que os algoritmos cometeram equívocos (ou classificações imprecisas) para analisar linguisticamente a razão pela qual há confusão na distinção das relações *Follow-up* e *Elaboration*, por exemplo. Sabe-se que esta última relação pode conter alguma informação temporal (ainda que não seja compreendida pela complementaridade), mas possui traços que não puderam ser sistematizados para a proposta de atributos (como conhecimento de mundo).

No cenário atual de identificação de relações CST de complementaridade, para o PB, tem-se o CSTParser (MAZIERO, PARDO, 2012). De acordo com os autores, a detecção das relações tem precisão aproximada de 70%, contando basicamente com informações sobre a redundância. Assim, acredita-se que os resultados dos algoritmos de classificação desenvolvidos para esse *parser* poderiam ser melhorados, uma vez que admitidos os atributos específicos de complementaridade.

Por fim, faz-se breve comentários sobre as hipóteses levantadas inicialmente sobre esta pesquisa:

a) a ideia de que "atributos superficiais e profundos de detecção da redundância são pertinentes para a identificação da complementaridade" ocorre, de fato, na superfície

linguística dos pares de sentenças que foram anotados com as relações de complementaridade. Entretanto, como demonstrado, durante a identificação do fenômeno em questão, essa hipótese não se confirmou;

b) a hipótese de "a complementaridade pode se manifestar na superfície linguística" foi confirmada, uma vez que essa manifestação pôde ser traduzida em atributos que ocorriam na materialidade do texto;

c) a ideia de que "métodos de detecção da complementaridade podem capturar os diferentes tipos de complemento" também foi confirmada, já que os atributos utilizados para identificação dos tipos de complemento partiram da proposta de organização tipológica das relações CST para o PB;

d) por fim, os atributos adotados nesta pesquisa também confirmam a hipótese de que "métodos de detecção da complementaridade capturam as relações CST" de complementaridade.

6.2 Limitações

Dada sua natureza, uma das limitações de se modelar o fenômeno da complementaridade em atributos, é que, por vezes, o fato de que ele se dá por inferências e/ou correferências. Diante da dificuldade se de modelar computacionalmente tais características, que dependem de conhecimento de mundo, por exemplo, acredita-se que parte do que caracteriza a complementaridade, sobretudo o tipo atemporal, não foi comtemplado. Com isso, a captura automática da complementaridade atemporal, fortemente marcada pela inferência, fica prejudicada.

Outra limitação desta esta pesquisa diz respeito ao estudo de *corpus*, especificamente, à a tarefa de se identificar os trechos das sentenças dos pares do CSTNews efetivamente envolvidos na complementaridade. Apesar de o *corpus* CSTNews ter sido anotado com base em um protocolo/manual de regras, os trechos das sentenças que levaram os anotadores a anotar as relações CST não estão explícitos e, por vezes, foi difícil recuperá-los pela análise dos pares. Muitas vezes, aliás, foi preciso consultar os textos-fonte na íntegra para compreender o porquê de certas sentenças terem sido anotadas com ums das relações CST de complementaridade.

Diante dessa dificuldade, não se descarta a possibilidade de que a especificação de valores referentes a alguns atributos ter sido equivocada.

6.3 Contribuições para a Linguística

Este trabalho contribui para a Linguística, em especial, à Linguística Descritiva, ao fornecer um levantamento das características do fenômeno da complementaridade entre sentenças advindas de textos distintos que abordam o mesmo assunto. Por meio deste trabalho, observou-se que não somente a redundância é uma das características da complementaridade, mas sim outros traços. Assim, pode-se dizer que se sabe mais sobre o referido fenômeno hoje do que antes do desenvolvimento desta pesquisa. Por fim, salienta-se que este trabalho propiciou a descrição de mais um fenômeno semântico multidocumento, pois, até então, apenas a redundância havia sido amplamente investigada no português e em outras línguas.

6.4 Contribuições para o PLN

As regras de identificação das relações CST e dos tipos de complementaridade têm o potencial de subsidiar métodos automáticos de detecção do fenômeno em questão, os quais são muito úteis para a SAM. Além de uma aplicação direta na SAM, sistemas de pergunta-resposta, simplificação textual, tradução automática e/ou recuperação de informação, dentre outros, podem ser beneficiados por regras de detecção automática da complementaridade.

Por exemplo, num sistema de pergunta-resposta, suponha-se que o usuário gostaria de saber o resultado das eleições presidenciais de 2014, no Brasil. O sistema pode buscar em bancos de dados distintos que abordem esse tema (jornais, por exemplo), e elaborar uma resposta eliminando a redundância que possa haver entre os textos-fonte, e identificando trechos que são complementares uns aos outros.

Ademais, os sistemas de identificação de relações CST podem passar por melhorias, já que agora contam com uma descrição e caracterização específica da complementaridade.

Por fim, esse trabalho caminha em tendências de descrições já realizadas para fenômenos linguísticos (mono ou multidocumento). Alguns trabalhos, como os de TABOADA e DAS (2013), TABOADA (2006, 2009), MAZIERO (2012), MAZIERO e PADO (2010), utilizam camadas de informações linguísticas para detecção de relações específicas, visando possíveis aplicações computacionais, inclusive a identificação automática desses fenômenos.

6.5 Trabalhos futuros

Como trabalhos futuros, ressalta-se a investigação do potencial das regras aqui aprendidas para a detecção automática de outras relações CST, como *Overlap*, que se caracteriza pela presença de informação complementar entre as sentenças de um par, sendo que há bastante redundância entre elas.

Outra possiblidade é modelar a complementaridade em outros atributos que, possivelmente, pautem-se em conhecimento de mudo (ainda que em pequena escala) e/ou em aspectos informacionais (“onde”, “quando” e “como”, por exemplo).

Além disso, é possível haver uma revisão da tipologia de organização das relações CST, proposta ao PB, já que algumas relações se comportam de maneira bastante similar (por exemplo, a relação *Overlap* em relação às demais relações de complementaridade). Após essa pesquisa, sabe-se que as relações CST de complementaridade não se comportam somente pela presença ou ausência de um aspecto temporal nas sentenças. Há casos, como demonstrado, que a complementaridade ocorre em função de outros fatos linguísticos ou extralinguísticos (como conhecimento de mundo e correferência).

REFERÊNCIAS

AFANTENOS, S. D.; DOURA, I.; KAPELLOU, E.; KARKALETSIS, V. Exploiting cross-document relations for multi-document evolving summarization. In: **Methods and Applications of Artificial Intelligence**, pp. 410-419. Springer Berlin Heidelberg. 2004.

ALEIXO, P. PARDO, T.A.S. **CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST.** Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 326. São Carlos-SP, 15p. 2008.

ALEIXO, P. PARDO, T.A.S. CSTTool: um parser multidocumento automático para o Português do Brasil. In: **Proceedings of the IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence – WTDIA**. Salvador, Bahia. 2008.

ALLAN, J. Automatic Hipertext Linking Type. In **Proceedings of Hypertext**. Washington D.C./USA. 1996.

BAPTISTA, J. HAGÈGE, C. MAMEDE, N. Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o HAREM II. In: **Actes de Encontros do Segundo HAREM**. 2008.

BICK, E. **The parsing system "PALAVRAS": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.** PhD Thesis. Aarhus Univesity. Denmark University Press. Outubro. p. 412. 2000.

CARDOSO, P.C.F. **Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo.** Tese de doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP. 2014.

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 Córpus Multidocumento CSTNews.** Série de Relatórios Técnicos do Instituto de

Ciências Matemáticas e de Computação, Universidade de São Paulo, n. 389. NILC-TR-12-07. São Carlos-SP, Junho, 18p 2012.

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

DAGAN, I. BENTIVOGLI, L. DANG, H.T. GIAMPICCOLO, D. MAGNINI, B. The Fifth PASCAL Recognizing Textual Entailment Challenge. In **Proceedings of Text Analysis Conference** (TAC'09). 2009.

FERREIRA, A. B. H. **Novo dicionário Aurélio**. Editora Nova Fronteira, 1999.

GASPERIN, C.V. LIMA, V.L.S**. Fundamentos do processamento estatístico da linguagem natural.** Série de Relatórios Técnicos da Faculdade de Informática – PUCRS. n. 021. Porto Alegre – RS. 2000.

HATZIVASSILOGLOU, J. L.; KLAVANS J.L.; HOLCOMBE, M. Simfinder: a flexible clustering tool for summarization. In: **Proccedings of NAACL Automatic Summarization Workshop**. Pittsburgh, PA, USA. 2001.

HOLT, R.C. **Very simple classification rules perform well on mosty commonly using dataset.** Machine Learning. 1993.

HOUAISS, A. VILLAR, M.S. **Dicionário Eletrônico Houaiss da Língua Portuguesa. S**ão Paulo. Objetiva, 2001.

KOCH, I.G.V. **Introdução à linguística textual.** São Paulo: Contexto. 2009.

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

LAGE, N. **Estrutura da notícia**. Ática, 2002.

MACCARTNEY, B. TROND, G. MARIE-CATHERINE, M. DANIEL, C. CHISTOPHER, D.M. Learning to recognize features of valid textual entailments. In: **Proceedings of HLT/NAACL.** 2006.

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

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 ACL'05 – Workshop on empirical modeling of semantic equivalence and entailment.** Ann Arbor-Michigan. Junho. P. 1-6. 2005.

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. p. 60-9. 2010.

MAZIERO, E.G. PARDO, T.A.S. DI FELIPPO, A. DIAS DA SILVA, B.C. A base de dados lexical e a interface web do TeP 2.0 – Thesaurus Eletrônico para o Português do Brasil. In: **Anais VI Workshop em tecnologia da informação e da linguagem humana.** Vila Velha. P.390-392. 2008.

MAZIERO, E.G.. **Identificação automática de relações multidocumento.** Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP. 2012.

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

MENEZES FILHO, L.A. PARDO, T.A.S. Detecção de Expressões Temporais no Contexto de Sumarização Automática. In: **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.

MIYABE, T. TAKAMURA, H. OKUMURA, M. Identifying a cross-document relation between sentence. In: **Proceedings of the Third International Joint Conference on Natural Language Processing**, v. 1, p. 141-148. 2008.

NEWMAN, P.S. JOHN C. B. Summarizing archived discussions: a beginning. **Proceedings of the 8th international conference on Intelligent user interfaces**. New York – USA. 2003.

QUILAN, R. **Programs for machine learning.** Morgan Kaufmann Publishers. San Mateo. 1993.

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,** v. 10, p. 74-83. 2000.

RADEV, D.R. MCKEOWN, K. Generating natural language summaries from multiple on-line sources. **Computational Linguistics**, v. 24, N. 3, pp. 469-500. 1998.

RADEV, D.R.; OTTERBACHER, J.; ZHANG, Z. CST Bank: A Corpus for the Study of Crossdocument Structural Relationships. In: **Proceedings of 4th International Conference on Language Resources and Evaluation.** 2004.

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. 2012.

SPARCK JONES, K. What might be in a summary?. **Information Retrieval**. v. 93, p. 9-26. 1993.

TABOADA, M. DAS, D. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. **Dialogue and Discourse**, v. 4, n. 2, pp. 249-281. 2013.

TABOADA, M. Implicit and explicit coherence relations. In J. Renkema (Ed.), **Discourse, of Course**. p. 127-140. Amsterdam and Philadelphia: John Benjamins. 2009.

TABOADA, M. **Building Coherence and Cohesion: Task-Oriented Dialogue in English and Spanish**. Amsterdam and Philadelphia: John Benjamins. 2004.

TAUFER, P. Massa de informações digitais pode ser usada em benefício da população. **Jornal da Globo**, 26.dez.2013. 2013. Disponível em: http://g1.globo.com/jornal-da-globo/noticia/2013/12/massa-de-informacoes-digitais-pode-ser-usada-em-beneficio-da-populacao.html. Acesso em: 02.02.2015.

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

TRIGG, R.; WEISER, M. TEXTNET: A Network-Based Approach to Text Handling. In **ACM Transactions on Office Information Systems**, v. 6. 1986.

WITTEN, I.H. FRANK, E. **Data mining: Practical machine learning tools and techniques.** Morgan Kaufmann Publishers. San Mateo. 2005.

WITTEN, I.H. FRANK, E. **Generating accurate rule sets without global optimization.** Working paper series. ISSN 1170-487X. 1998.

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.