Programa de Pós-Graduação em Linguística

PPGL

APROFUNDAMENTO DA CARACTERIZAÇÃO LINGUÍSTICO-COMPUTACIONAL DA COMPLEMENTARIDADE EM UM CORPUS JORNALÍSTICO MULTIDOCUMENTO

Jackson Wilke da Cruz Souza

SÃO CARLOS
2019

Universidade Federal de São Carlos

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS

PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

APROFUNDAMENTO DA CARACTERIZAÇÃO LINGUÍSTICO-COMPUTACIONAL DA COMPLEMENTARIDADE EM UM CORPUS JORNALÍSTICO MULTIDOCUMENTO

JACKSON WILKE DA CRUZ SOUZA

BOLSISTA CAPES

Tese apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Linguística, área de concentração: Descrição, Análise e Processamento Automático de Línguas Naturais.

Orientadora: Profa. Dra. Ariani Di Felippo.

São Carlos - SP

Fevereiro – 2019

SOUZA, Jackson Wilke da Cruz

Aprofundamento da caracterização linguístico-computacional da complementaridade em um corpus jornalístico multidocumento / Jackson Wilke da Cruz SOUZA. -- 2019.
217 f. : 30 cm.

Tese (doutorado)-Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador: Ariani Di Felippo
Banca examinadora: Oto Araújo Vale, Thiago Alexandre Salgueiro Pardo, Paula Christina Figueira Cardoso, Erick Galani Maziero
Bibliografia

1. Sumarização Automática Multidocumento. 2. Processamento Automático de Línguas Naturais. 3. Descrição linguística. I. Orientador. II. Universidade Federal de São Carlos. III. Título.

Ficha catalográfica elaborada pelo Programa de Geração Automática da Secretaria Geral de Informática (SIn).
DADOS FORNECIDOS PELO(A) AUTOR(A)
Bibliotecário(a) Responsável: Ronildo Santos Prado – CRB/8 7325

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Linguística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado do candidato Jackson Wilke da Cruz Souza, realizada em 27/02/2019:

Profa. Dra. Ariani Di Felippo
UFSCar

Prof. Dr. Oto Araujo Vale
UFSCar

Prof. Dr. Thiago Alexandre Salgueiro Pardo
USP

Profa. Dra. Paula Christina Figueira Cardoso
UFLA

Prof. Dr. Erick Galani Maziero
UFLA

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Erick Galani Maziero e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Profa. Dra. Ariani Di Felippo

AGRADECIMENTOS

Ufa! O caminho foi longo até aqui! Foram tantas experiências, tantas pro/privações, foram muitos *nãos*, mas os *sins* que me acenderam nesse percurso de doutoramento valeram muito a pena por seus pesos e importâncias. Certamente foi uma aventura inesquecível: para além das dificuldades, sou muito grato por tudo que aconteceu até aqui, pois sou fruto não da proposta, mas do caminho, do percurso.

Agradeço, então, a Deus: a Pessoa mais importante que conheci (*e continuo conhecendo*) nesse caminho que escolhi (*?*) viver; quem me conhece por dentro e por fora, quem sempre se antecipou às lágrimas e alegrias. Te agradeço, *Éternel*.

Pode parecer difícil acreditar que Deus pode abraçar, mas Ele pode o fazer, encarnando Seu amor e cuidado em pessoas. Assim, agradeço esse amor encarnado em minha família: Marlene, José e Jéssica. Como vocês foram importantes nesse caminho todo. Foram dez (*longuíssimos*) anos, foram mais de 150.000km rodados para nos encontrarmos aos fins de semana, em aniversários e feriados. Obrigado por me atenderem de madrugada, por me ouvirem, por me encorajarem, por me suportarem (*em seu sentido mais polissêmico possível!*). Obrigado por tudo que abriram mão para que esse sonho de formação intelectual fosse possível. Obrigado pela coragem baiana que nos motivou a sair do sertão (*me espere, que eu volto!*) e chegarmos à selva de pedra paulista, sem estrutura, sem conhecer nada e ninguém, e chegarmos JUNTOS até aqui: vocês se doutoraram juntamente comigo. Obrigado, família!

Obrigado, Ariani Di Felippo, minha orientadora. Esse caminho foi árduo e tortuoso, às vezes, mas com bastante e incessante aprendizado. Tive tantas dúvidas, tantas perguntas (*continuo tendo muitas delas*), mas você sempre esteve disponível e acessível: a jornada foi mais interessante e encorajadora sob sua orientação. Que novos desafios venham por aí!

Obrigado aos professores que compuseram a banca até aqui. Profes. Drs. Paula, Oto, Erick e Thiago. Obrigado pelo tempo que dispuseram com a leitura atenta a meu texto e a minhas ideias. Que nossos caminhos acadêmicos e profissionais se cruzem mais e mais vezes!

Agradeço também aos meus amigos, de perto e de longe, os já voltaram ao Lar e aos que ainda estão caminhando por aqui. Obrigado por me amarem mesmo eu demorando uma vida para responder no *whatsapp*! Obrigado por respeitarem meus silêncios, minhas distâncias, em especial na reta final do doutoramento. Há tanto a agradecer e há tantos nomes que me atravessam nesse momento, mas espero, de verdade, que você, meu amigo e minha amiga, ao ler esse trecho se

enxergue aqui e sinta que meu coração transborda de alegria em ter você como companheiro e companheira de jornada. Espero ainda que possamos tomar aquele café demorado, sairmos para dançar um forró ou uma dança latina; possamos parar somente para rirmos um do outro, lembrarmos de todas as histórias que vivemos juntos, falarmos como as coisas mudaram... Espero retribuir todo o amor e carinho que sinto e recebo por ser seu amigo.

Obrigado aos amigos de projeto, trabalhos e formação profissional. Foram atividades que não estavam ligadas diretamente à universidade, mas que foram basilares ao processo de doutoramento. Muitos dos sonhos que tenho nasceram ou se solidificaram a partir de nossas conversas, idas a praças, abrigos, orfanatos, prisões e ao sertão nordestino. Obrigado por cada abraço e encorajamento; por cada oportunidade de me reconhecer e me encontrar em cada criança e adolescentes que tocamos durante todos esses anos: vocês deixaram suas digitais em mim, alguns sem nem precisar tocar fisicamente.

Sou grato à UFSCar em seus diversos e múltiplos meios. Entrei aqui e foram muitas lições aprendidas, dentro e fora do espaço ideológico de formação acadêmica. Mas saio hoje com a lição mais importante aprendida: há dez anos, moreno, hoje, NEGRO! Obrigado por acender essa consciência em mim, além de ter aberto portas e apontado caminhos. Agradeço a cada professor (*por cada aula, conversa, cafés!*), por cada funcionário (*aqueles que ainda estão no kronos e os que já estão no kairós*), cada aluno, cada participante de projetos: vocês foram essenciais nesse tempo por aqui.

Obrigado ao PPGL (alunos, professores e funcionários) por ter abraçado muitas ideias (eventos, minicursos, rodas de conversas...) que tive no meio do caminho de doutoramento: tudo isso foi muito importante para minha formação acadêmica, profissional e pessoal! Um agradecimento especial à Profa. Dra. Gladis M. B. Almeida por ter-me concedido espaço em seu laboratório (*mesmo eu não trabalhando diretamente com terminologia!*) e por amizade, parceria e compreensão durante minha carreira acadêmica na UFSCar, em especial, neste doutoramento: levarei marcas importantes suas em mim. Muito obrigado também ao professor Jorge Baptista por ter concedido algumas horas de seu tempo em uma de suas visitas à UFSCar: aquela troca de ideias reacendeu o questionador científico dentro de mim e tenho certeza que o fim deste doutoramento poderia ser outro (um tanto mais trabalhoso e menos investigativo) se não tivéssemos aquela conversa.

Obrigado a Capes pelo financiamento deste projeto.

Digo: o real não está na saída nem na chegada: ele se dispõe para a gente é no meio da travessia.
Guimarães Rosa

RESUMO

No contexto de disseminação da informação digital, projeta-se que 3.3 Zettabytes de informação estarão em circulação na Web em 2021. Neste contexto, subáreas do Processamento Automático de Línguas Naturais (PLN) desenvolvem soluções linguístico-computacionais para dinamizar o pouco tempo que o usuário tem frente à demanda de informações, como a Sumarização Automática Multidocumento (SAM), em que se visa criar sumários automáticos a partir de coleções de textos-fonte que versam sobre um mesmo assunto. Com o objetivo de viabilizar a SAM e o aperfeiçoamento da seleção de conteúdo dos sumários, algumas pesquisas da área realizaram descrições linguísticas de fenômenos multidocumentos. Um desses fenômenos é a Complementaridade, a qual ocorre quando, em um par de sentença (S1,S2), S2 elabora alguma informação apresentada por S1. O modelo teórico *Cross-Document Structure Theory* (CST) traduz essa complementaridade em três relações semânticas: *Historical Background* e *Follow-up* (temporal) e *Elaboration* (atemporal). Sabe-se que atributos linguísticos que denotam informações temporais são relevantes para identificar automaticamente tais relações CST, obtendo classificadores automáticos com 75% de precisão. Assim, sob a hipótese de que informações linguísticas profundas pudessem gerar classificadores mais eficientes, propôs-se um conjunto mais refinado de atributos que caracterizam o fenômeno em questão. Após a análise manual dos pares de sentenças anotados com as relações CST de complementaridade do *corpus* CSTNews, chegou-se a uma tipologia de 32 sinalizadores, organizados em anáfora, estrutura textual, morfologia, sintaxe, semântica, pragmática. Com o auxílio de algoritmos simbólicos de Aprendizado de Máquina, foi possível construir e treinar novos classificadores, cuja precisão superou o estado-da-arte. Assim, contribui-se com (i) a Linguística Descritiva, já que sistematicamente apresenta-se um conjunto de sinalizadores, organizados tipologicamente, que evidenciam e caracterizam a complementaridade em pares de sentenças de textos jornalísticos e com o (ii) PLN, pois produziu-se uma descrição mais refinada e específica para a identificação automática da Complementaridade e, consequentemente, o aprimoramento de seleção de conteúdo para os sumários automáticos.

Palavras-chave: Descrição linguística. Complementaridade. Fenômenos multidocumento. Sumarização Automática Multidocumento. Processamento Automático de Línguas Naturais. Análise de texto.

ABSTRACT

In the context of the dissemination of digital information, CISCO, an agency of web security, projects that 3.3 Zettabytes of information will be circulated on the Web in 2021. In this context, sub-areas of Automatic Natural Languages Processing (NLP) develop linguistic-computational solutions to dynamize the short time the user has in front of the demand of information in circulation on web. One of these sub-areas is Automatic Multi-document Summarization (AMS), which aims to create automatic summaries from collections of source texts that deal with the same subject. In order to make possible the selection of contents to automatic summaries and to improve this technique, some studies are based in linguistic descriptions of multi-document phenomena. One of these phenomena is complementarity, which occurs when, in a sentence pair (S1, S2), S2 elaborates some information presents in S1. The theoretical model Cross-Document Structure Theory (CST) translates the complementarity into three semantic relations: Historical Background and Follow-up (temporal) and Elaboration (timeless). Some studies in this area indicate that (superficial) linguistic temporal attributes are relevant to automatically identify such CST relations, obtaining automatic classifiers with 75% accuracy. Thus, under the hypothesis that deep linguistic information could generate more efficient classifiers, we propose a refined set of attributes that characterize the complementarity. After the manual analysis of the pairs of sentences annotated with the CST relations of complementarity of CSTNews *corpus*, we built a typology of 32 signs, organized in seven categories, namely: anaphora, textual structure, morphology, syntax, semantics, pragmatics. Using symbolic algorithms of Machine Learning, it was possible to construct and train new classifiers, whose accuracy surpassed the state-of-the-art. Thus, we contribute with (i) Descriptive Linguistics, as a typology organized in signs that present systematically the evidences and characteristics of complementarity in sentences pairs of journalistic texts, and with (ii) NLP, as it produced a more refined and specific description for the automatic identification of complementarity and consequently the selection of content to the automatic multi-document summaries.

Keywords: Language description. Complementarity. Multi-document phenomena. Automatic Multi-document Summarization. Automatic Natural Language Processing. Text analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Neste capítulo introdutório, viso delimitar o tema de meu trabalho, além de apresentar minhas motivação e contextualização para o desenvolvimento desta pesquisa no cenário do PLN, bem como apresento meus objetivos e hipóteses traçados. Por fim, defino a metodologia que abordei, a qual estruturou este estudo.

1.1 Motivação e Contextualização

A disponibilização, circulação e acesso à informação digital têm aumentado exponencialmente nos últimos anos. Um relatório[1] publicado pela agência Cisco-Visual-Networking-Index[2] em 2017 aponta que foi produzido 1 Zettabyte de informação na Web em 2016. As projeções para 2021 mostram que serão produzidos mais de 3,3 Zettabytes.

Diante desse cenário, subáreas do Processamento Automático de Línguas Naturais (PLN) podem produzir soluções computacionais para lidar com essa grande quantidade de dados disponível ao usuário. Subáreas que lidam com produção e seleção de conteúdo são aquelas que têm tido destaque nos últimos anos, produzindo, por exemplo, sistemas de análise de sentimento, pergunta e resposta e sumarizadores automáticos.

Em especial, na Sumarização Automática objetiva-se desenvolver sistemas automáticos que possam produzir automaticamente sumários (resumos) com base em texto-fonte que versam sobre um mesmo assunto. Quando a sumarização é feita a partir de um único texto-fonte, diz-se que houve Sumarização Automática Monodocumento; quando tal processo é

[1] Disponível em: https://goo.gl/KY74RH. Acesso em: 23/04/2018.
[2] Disponível em: https://goo.gl/MA6xr1

realizado com base em dois ou mais textos-fonte, diz-se que houve Sumarização Automática Multidocumento (doravante, SAM) (RADEV, 2000).

A produção de sumários automáticos multidocumento tem sido motivada pela relação que há entre a imensa quantidade de informação disponível aos usuários e o pouco tempo que estes têm para assimilá-la (MANI, 2001). Tais sumários são comumente *extratos informativos*[3] que dispensam a leitura integral dos textos-fonte e *genéricos* (isto é, sumários que não visam um público-alvo específico), ou seja, sumários compostos por sentenças extraídas integralmente dos textos-fonte que veiculam o conteúdo central da coleção (KUMAR *et al.*, 2012).

Os métodos utilizados para a produção dos sumários extrativos são tidos como superficiais, pois utilizam pouco ou nenhum conhecimento linguístico, baseando-se em pistas na superfície textual ou estatísticas. Assim, geram-se, por um lado, extratos com robustez, escalabilidade e baixo custo de desenvolvimento e, por outro, com menos coerência, coesão e informatividade.

Diante desse cenário, alguns trabalhos têm proposto métodos profundos que utilizam conhecimento linguístico codificado em gramáticas, repositórios semânticos ou modelos de discurso. O desenvolvimento desses métodos/sistemas[4] é custoso e sua aplicação é mais restrita, ainda que o desempenho seja superior, pois resulta em sumários mais coerentes, coesos e informativos. Tais métodos/sistemas, aliás, podem gerar extratos ou *abstracts* (isto é, produzidos pela reescrita dos textos-fonte).

Tendo em vista a produção de um sumário extrativo, informativo e genérico, é preciso selecionar as sentenças mais importantes de uma coleção de textos-fonte, evitando-se que elas sejam redundantes e contraditórias entre si, mas permitindo que sejam complementares. Em outras palavras, há a necessidade de identificar os fenômenos linguísticos que decorrem da análise multidocumento, que são *redundância*, *complementaridade* e *contradição*. Diz-se isso porque (i) as sentenças mais redundantes na coleção veiculam suas principais informações e,

[3] Por uma questão de *layout*, optou-se por destacar, no corpo de texto, palavras e/ou expressões importantes ou estrangeiras por meio de itálico.

[4] De acordo com Sparck Jones (1993), a arquitetura dos os métodos/sistemas de sumarização automática engloba três fases: (i) *análise*, em que os textos-fonte são interpretados, extraindo-se uma representação formal dos mesmos; (ii) *transformação*, em que o conteúdo dos textos-fonte é condensado em uma representação interna do sumário, a qual é resultante da seleção de conteúdo; (iii) *síntese*, em que o sumário é produzido em língua natural a partir do conteúdo selecionado.

por isso, seu conteúdo deve constar no sumário, (ii) as sentenças selecionadas que veiculam podem compor o sumário e (iii) as sentenças redundantes ou contraditórias entre si não devem ser selecionadas para o sumário.

Assim, é preciso identificar, na fase de análise[5], os fenômenos de conteúdo típicos da multiplicidade de textos-fonte, os quais estão ilustrados pelos pares de sentenças (S1 e S2) em (1), (2) e (3), respectivamente, extraídos de textos-fonte que relatam um acidente aéreo em Bukavu, no Congo.

(1)

S1: 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.

S2: Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu.

(2)

S1: Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.

S2: O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.

(3)

S1: Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu.

S2: A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.

Entre as sentenças de (1), por exemplo, há uma relação de *redundância*, já que, por serem basicamente idênticas, as sentenças expressam o mesmo conteúdo, no caso, as informações de que o avião caiu e não houve sobreviventes. Entre as sentenças de (2), por sua vez, há uma relação de *complementaridade*, já que a S2 acrescenta informação à S1, em que a segunda especifica que o avião "explodiu e se incendiou". E, finalmente, entre as sentenças de (3),

[5] Em um método/sistema de SAM que se baseia em conhecimento discursivo, a análise é feita por um analisador automático discursivo, como o CSTParser (MAZIERO; PARDO, 2011), desenvolvido para o Português do Brasil (PB).

observa-se uma relação de *contradição*, já que a S1 informa que a causa da queda do avião foi o mau tempo, enquanto a S2 aponta que foi um choque contra uma montanha.

Na literatura, encontram-se diversas propostas de análise automática de segmentos textuais advindos de documentos distintos que abordam o mesmo tema. As propostas mais proeminentes têm se baseado na detecção de relações (de sentido) entre sentenças analisadas em pares. As relações mais investigadas são do modelo *Cross-document Structure Theory* (CST) (RADEV, 2000).

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

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

Quadro 1: Conjunto refinado de relações CST de Maziero *et al.* (2010).
Fonte: Maziero *et al.* (2010).

De acordo com a tipologia proposta por Maziero *et al.* (2010) (cf. pág.25), as relações CST que capturam a complementaridade entre sentenças de um par (S1 e S2) advindas de textos-fonte distintos manifestam-se 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 presente em S1.

Além disso, Maziero *et al.* (2010) organizam as relações de complementaridade em temporais ou atemporais. As relações temporais podem ser de dois tipos diferentes. Dado um par de sentenças, S1 e S2, as mesmas são complementares do subtipo temporal quando (i) S2 apresenta informações históricas/ passadas sobre algum elemento presente em S1 (no modelo CST, essa relação é rotulada como *Historical Background*) ou (ii) S2 apresenta acontecimentos/eventos que sucederam os acontecimentos/eventos presentes em S1, sendo

que os acontecimentos em S1 e em S2 devem ser relacionados e devem ter ocorrido em um intervalo curto de tempo (no modelo CST, essa relação é rotulada como *Follow-up*).

A relação CST atemporal não ocorre entre um evento que sucede ou procede outro evento. Ela se estabelece quando, dado um par de sentenças, S1 e S2, S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1. Além disso, o elemento elaborado em S2 deve ser o foco de S2. No modelo CST, essa relação é rotulada por *Elaboration.*

Segundo Zhang e Radev (2005), as relações CST somente podem ocorrer entre sentenças que possuem algum tipo de sobreposição de conteúdo e/ou forma. Por essa razão, a identificação automática das relações CST de conteúdo (inclusive a complementaridade) tem sido feita quase que exclusivamente com base na similaridade lexical existente entre as sentenças. A similaridade pode ser modelada por um conjunto de atributos linguísticos (p.ex.: sobreposição de palavras de conteúdo) e capturada por várias medidas estatísticas (p.ex.: *word overlap*) que, mediante o valor obtido, indicam o fenômeno (redundância, complementaridade ou contradição) e as relações CST correspondentes (p.ex.: ZHANG *et al.*, 2002, 2003; ZHANG; RADEV, 2005; MAZIERO *et al.*, 2010).

Para o Português do Brasil (PB), o CSTParser é capaz de identificar as relações CST com precisão aproximada de 70%, baseando-se em atributos similares aos de Zhang *et al.* (2002, 2003) e Zhang e Radev (2005). 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, dentre outros (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, dentre outros. (HATZIVASSILOGLOU *et al.*, 1999, 2001; KUMAR *et al.*, 2012; SOUZA, 2015; SOUZA; DI-FELIPPO, 2018).

Do que foi exposto, observa-se que as relações CST de conteúdo, inclusive as de complementaridade, são identificadas em função de alguns atributos linguísticos que

capturam a similaridade entre duas sentenças, posto que sentenças complementares apresentam certo conteúdo redundante.

A identificação da complementaridade foi inicialmente estudada por Souza (2015). De acordo com o autor, tal fenômeno pode ser identificado com base em informações linguístico-estruturais, capturadas por atributos que evidenciem a complementação temporal entre as sentenças de um par. Objetivando desenvolver métodos específicos para a identificação automática dos tipos de complementaridade (temporal e atemporal) e as relações que os codificam (*Historical background*, *Follow-up* e *Elaboration*), o autor utilizou algoritmos de Aprendizado de Máquina (AM) que, com base em um aprendizado supervisionado, geraram classificadores que melhor distinguem a relação *Historical background* de *Elaboration*. A identificação da relação *Follow-up* ainda apresentava equívocos, a ponto de os classificadores confundirem-na com a relação *Elaboration*.

Dessa forma, dando continuidade ao trabalho realizado por Souza (2015), foram propostos os objetivos descritos na próxima subseção.

1.2 Objetivos e Hipóteses

O objetivo deste trabalho foi aprofundar a descrição da complementaridade com vistas a contribuir para a Linguística Descritiva e o PLN. De acordo com Taboada e Das (2013), investigar o comportamento de relações de sentido é uma das questões centrais para diversas aplicações em PLN, tal como a SAM. Os autores propõem que a identificação e a classificação das relações devem se basear na detecção de pistas linguísticas ou estruturais que caracterizam as relações a serem estudadas. Além disso, Taboada e Das (2013) afirmam que tais pistas sempre são reconhecidas por humanos, mas nem sempre podem ser traduzidas em atributos detectáveis por máquina.

Com base nisso, objetivou-se descrever o fenômeno da complementaridade com base em um *corpus* anotado com as relações CST, a fim de identificar *sinais* que caracterizam a complementaridade e organizá-los tipologicamente segundo sua natureza linguística, bem como propor uma modelagem dos mesmos para a detecção automática do fenômeno linguístico em questão.

Especificamente, objetivou-se:

a) produzir uma descrição ampla e refinada em textos jornalísticos, observando as características específicas do fenômeno e das relações CST que o traduzem;

b) identificar as características linguístico-estruturais capazes de distinguir os diferentes tipos de complementaridade.

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

- **Hipótese 1:** Uma vez que as relações que traduzem a complementaridade são relações de sentido, há evidências textuais nas quais os falantes de um idioma (neste caso, os anotadores do *corpus*) se pautam para identificá-las;
- **Hipótese 2:** A complementaridade se manifesta por sinais que, ora são textualmente marcados, ora dependem de conhecimento extra-linguístico.

Tendo em vista o cumprimento dos objetivos aqui descritos, estabeleceram-se as etapas metodológicas, a seguir.

1.3 Metodologia

Equacionou-se metodologicamente esta pesquisa em 6 tarefas.

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

Tarefa 2 – Seleção e recorte do *corpus*: consistiu na seleção do *corpus* multidocumento CSTNews, que é composto por textos advindos de fontes jornalísticas distintas. Além disso, englobou um recorte no *corpus*, delimitando (i) uma parcela para estudo da complementaridade e investigação de sinalizadores, e (ii) uma parcela para teste e proposição de algoritmos (automáticos).

Tarefa 3 – Análise e caracterização do *corpus*: consistiu na análise e caracterização manual da parcela de estudo do *corpus* construída na Tarefa 2. Especificamente, os pares cujas sentenças estavam conectadas pelas relações CST de complementaridade foram

analisados com o objetivo de identificar dispositivos de sinalização do fenômeno. Como resultado, obteve-se um conjunto de sinalizadores que foram utilizados na caracterização dos pares anotados com complementaridade.

Tarefa 4 – Proposição de tipologia de atributos da complementaridade: consistiu em organizar as características da complementaridade em dois grandes grupos, tratável e não-tratável computacionalmente, afim de identificar possíveis atributos que pudessem ser relevantes para a identificação (automática) da complementaridade.

Tarefa 5 – Identificação de métodos de detecção das relações CST de complementaridade: consistiu no estudo semiautomático da correlação entre os métodos já identificados na Tarefa 4 e as relações CST de complementaridade (*Historical Background, Follow-up* e *Elabaration*).

Tarefa 6 – Avaliação e extração de conhecimento: consistiu na aplicação dos métodos mais eficientes identificados nas Tarefas 4 e 5 à parcela de avaliação do *corpus*. Objetivou-se, nesta tarefa, avaliar a pertinência dos métodos na tarefa de identificar as relações CST de complementaridade por meio de algoritmos de aprendizado de máquina.

1.4 Estrutura da Tese

Esta tese está organizada em 5 seções. No Capítulo 2, apresenta-se a revisão da literatura. Como resultado dessa revisão, apresenta-se o modelo CST e as suas relações de complementaridade e descrevem-se os principais métodos para a identificação automática das relações CST. No Capítulo 3, descrevem-se a seleção e estudo do *corpus* CSTNews. No Capítulo 4, apresenta-se a tipologia de sinais genéricos e específicos que caracterizam a complementaridade. No Capítulo 5, apresentam-se os testes e a avaliação da identificação automática das relações CST de complementaridade. No Capítulo 6, tecem-se considerações finais sobre esta tese.

Capítulo 2

REVISÃO DA LITERATURA

Neste capítulo, apresento a Revisão da Literatura, em que busquei conceitos relevantes à descrição de fenômenos linguísticos multidocumento, a fim de fomentar um arcabouço teórico-metodológico para o desenvolvimento deste trabalho. Além disso, apresento o modelo teórico CST, e destaco as relações discursivas que o compõem. Por fim, faço ainda algumas considerações sobre Gênero Textual, já que o objetivo do trabalho é estudar as relações CST em textos jornalísticos do PB.

2.1. O modelo *Cross-document Structure Theory*

Radev (2000), ao propor o modelo semântico *Cross-document Structure Theory* (CST), objetivava tratar fenômenos linguísticos intertextuais com a finalidade de gerar sumários de maneira automática. O autor aponta que, ao analisar informações intertextuais, fenômenos linguísticos acontecem por haver sobreposição, contradição e complementação de informações, por conta da natureza da tarefa. Como resultado, o modelo semântico proposto visa apresentar rótulos aos fenômenos linguísticos, organizando, em pares, as unidades (p.ex. sentenças) intertextuais em que tais fenômenos se manifestam, como exemplificado no Quadro 2.

Texto I

1. Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo, matou 17 pessoas na quinta-feira à tarde, informou hoje um porta-voz das Nações Unidas.
2. As vítimas do acidente foram 14 passageiros e três membros da tripulação.
3. Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu.

Texto II

1. Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.
2. As vítimas do acidente foram 14 passageiros e três membros da tripulação.
3. Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu.

Texto III

1. Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.
2. Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.
3. A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.
4. Acidentes aéreos são frequentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.

Quadro 2: Textos sobre um acidente aéreo no Congo
Fonte: Elaborado pelo autor.

No Quadro 2, ilustra-se um conjunto de textos que abordam um mesmo tópico, no caso, um acidente aéreo em Bukavu, no Congo. Com base no referido quadro, discorre-se sobre os diferentes fenômenos observados por Radev (2000). Os três fragmentos textuais são provenientes de textos jornalísticos em Português. Os Textos I e II são bastante similares, variando apenas em parcas ocorrências lexicais e construções frásticas. Dessa comparação inicial, aponta-se que há redundância (praticamente total) entre as Sentenças 1 de ambos os textos.

O Texto III, em relação aos Texto I e II, possui mais variação de forma, mas ainda contém sobreposição de conteúdo e/ou informação em relação aos demais textos. Como exemplo disso tem-se que a Sentença 4 do Texto III pode ser considerada complementar à Sentença 1 do Texto I e mesmo à Sentença 1 do Texto II, já que apresenta um fato anterior, de aspecto habitual relacionado ao assunto do conjunto de textos.

Ao realizar análises intertextuais semelhantes a ilustrada no Quadro 2, Radev (2000) propôs um conjunto de rótulos com base nas relações *Rhetorical Structure Theory* (RST)[6]

[6] Mann e Thompson (1987) objetivaram analisar textos por meio da geração de árvores com unidades de conteúdo (palavras, sentenças ou mesmo parágrafos) que estejam relacionadas por alguma relação. Caso todas as unidades de conteúdo do texto estejam conectadas entre si, tem-se um texto coeso e coerente e, concomitantemente, com um nível informacional relevante. Por exemplo, no Quadro 2, a Sentença 1 do Texto III desenvolve o evento principal narrado, atribuindo-lhe detalhes e, assim, recebe o rótulo *Attribution*; já a

(MANN; THOMPSON, 1987), o qual é capaz de representar como as informações estão organizadas textualmente, evidenciando fatores de coesão e coerência. De acordo com a teoria, se um texto é bem construído (informacional e estruturalmente), do ponto de vista da coerência e da coesão, suas unidades discursivas (comumente, proposições ou sentenças) estarão conectadas, resultando em uma espécie de estrutura retórica.

A proposta do autor do modelo CST não é construir uma estrutura retórica intratextual, mas conectar ou relacionar as unidades discursivas intertextualmente, sendo que tais conexões evidenciam os fenômenos linguísticos. Assim, ainda que se baseie na RST, o modelo proposto por Radev (2000) não é retórico, mas semântico.

Além da RST, Radev (2000) inspira-se em um de seus trabalhos anteriores (RADEV; MCKEOWN, 1998), em que, ao analisar conjuntos de textos sobre um mesmo assunto, fez observações relevantes a respeito do agrupamento de informações redundantes.

Como resultado do estudo, Radev (2000) apresenta um conjunto de relações semânticas que rotulam fenômenos linguísticos (a saber, similaridade, contradição, variação de escrita e complementaridade) que se estabelecem intertextualmente. Na proposta original, a CST compreende um conjunto de 24 relações, as quais estão descritas no Quadro 3.

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*

Quadro 3: Conjunto original de relações CST.
Fonte: Radev (2000)

De acordo com Radev (2000), tais relações podem ocorrer nos níveis lexical, sintagmático, sentencial e intertextual. Em outras palavras, as relações CST podem rotular conexões

Sentença 4 do mesmo texto, serve como detalhamento histórico sobre os acidentes aéreos na região (em que um deles é topicalizado pela Sentença 1) e, assim, recebe o rótulo *Elaboration*.

estabelecidas entre unidades informacionais desses diferentes níveis linguísticos. Na Figura 1, ilustram-se os diferentes níveis em que as relações CST podem ser identificadas.

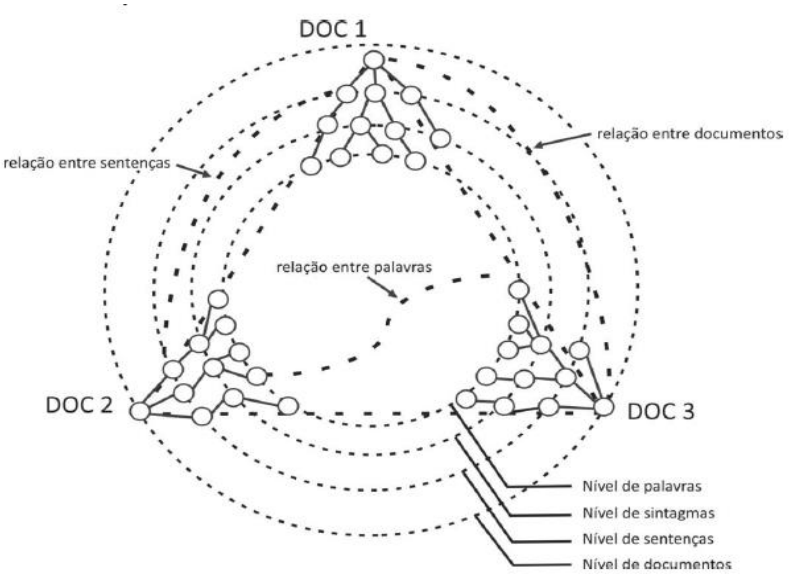

Figura 1: Esquema de relacionamento CST
Fonte: Radev (2000).

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

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

Tal como na RST, a identificação de uma relação CST está sujeita a ambiguidades (AFANTENOS *et al.*, 2004; ZHANG *et al.*, 2002, 2003), já que pode haver mais de uma

relação possível entre os segmentos textuais. Assim, com o objetivo de reduzir essa ambiguidade, alguns trabalhos modificaram o conjunto original de Radev (2000).

Zhang *et al.* (2002, 2003) realizaram análises de *corpus* em inglês e reduziram o conjunto original para 18 rótulos, 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*).

Ao realizar a análise de textos jornalísticos, Aleixo e Pardo (2008a) observaram que algumas relações originais não ocorriam no *corpus*. Como resultado, alguns rótulos foram excluídos ou unificados por serem considerados muito similares, ajustando o conjunto a 14 rótulos. A partir do refinamento de Aleixo e Pardo (2008a), Maziero *et al.* (2010) elaboram uma tipologia para as relações CST, ilustrada pela Figura 2.

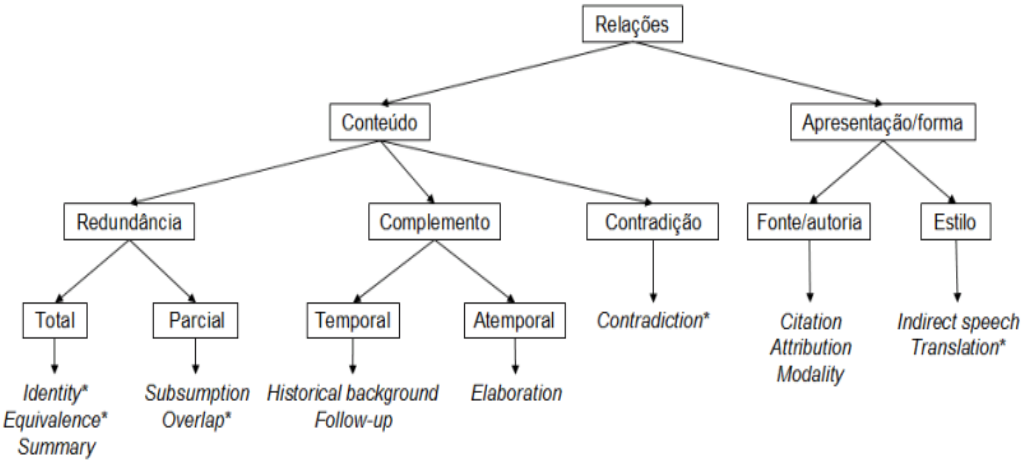

Figura 2: Tipologia de relações CST para o PB.
Fonte: Maziero *et al.* (2010).

Nessa tipologia ilustrada na Figura 2, as relações CST foram organizadas em dois grandes grupos: (i) relações de conteúdo (isto é, que rotulam os relacionamentos semânticos entre sentenças) e (ii) relações de forma (ou seja, que rotulam relacionamentos entre sentenças com base na forma). Cada grupo apresenta subdivisões. As relações de conteúdo podem ser classificadas nas categorias "redundância", "complemento" e "contradição". As relações da categoria "redundância", em especial, podem ser parciais ou totais, e as da categoria "complemento" podem ser temporais ou atemporais. As relações de forma, por sua vez,

podem ser do tipo "fonte/autoria" ou "estilo". Na Figura 2, o símbolo (*) indica que a relação não tem direcionalidade.

A partir dessa tipologia, Maziero *et al.* (2010) definiram cada uma das 14 relações CST. A definição das relações engloba (i) nome (ou rótulo), (ii) tipo, (iii) direcionalidade, (iv) restrição e (v) comentários (quando pertinente), ilustrados no Quadro 4.

RELAÇÃO	TIPO	DIR.	RESTRIÇÕES	COMENTÁRIOS
Identity	Conteúdo→ Redundância Total	Nula	As sentenças devem ser idênticas	---
Equivalence	Conteúdo→ Redundância Total	Nula	As sentenças apresentam o mesmo conteúdo, mas expresso de forma diferente.	---
Summary	Conteúdo→ Redundância Total	S1 ← S2	S2 apresenta o mesmo conteúdo que S1, mas de forma mais compacta.	*Summary* é um tipo de *Equivalence*, mas *Summary* deve haver diferença significativa de tamanho entre as sentenças.
Subsumption	Conteúdo→ Redundância Parcial	S1 → S2	S1 apresenta as informações contidas em S2 e informações adicionais.	S1 contém X e Y, S2 contém X.
Overlap	Conteúdo→ Redundância Parcial	Nula	S1 e S2 apresentam informações em comum e ambas apresentam informações adicionais distintas entre si.	S1 contém X e Y, S2 contém X e Z.
Historical background	Conteúdo → Complemento Temporal	S1 ← S2	S2 apresenta informações históricas sobre algum elemento presente em S1.	O elemento explorado em S2 deve ser o foco de S2; se forem apresentadas informações repetidas, considere outra relação (p.ex.: *Overlap*); se os eventos em S1 e S2 forem relacionados, pondere sobre a relação *Follow-up.*
Follow-up	Conteúdo → Complemento Temporal	S1 ← S2	S2 apresenta acontecimentos que acontecem após os acontecimentos em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si.	---
Elaboration	Conteúdo →	S1 ← S2	S2 detalha/refina/elabora	O elemento elaborado

	Complemento Atemporal		algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1.	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.	---

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

Como mencionado, o conjunto de 14 relações de Maziero *et al.* (2010) foi proposto a partir da anotação manual do *corpus* CSTNews (CARDOSO *et al.*, 2011). Na Tabela 1, a ocorrência dos fenômenos multidocumento de conteúdo, apenas, e suas respectivas relações CST no referido *corpus* são apresentados.

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

Tabela 1: Relação quantitativa das relações CST.
Fonte: Maziero *et al.* (2010).

Com base na Tabela 1, observa-se que a redundância e a complementaridade são os fenômenos multidocumento mais frequentes no *corpus*, bem como as relações CST que os codificam. Isso pode ser justificado pelo fato de os fenômenos serem identificados entre textos que abordam um mesmo assunto. Tais textos-fonte são do gênero textual "notícia", cujas características serão descritadas na subseção 3.1.2.

A seguir, descreve-se especificamente a complementaridade com base na definição proposta por Maziero *et al.* (2010), e se exemplificam as relações CST que a codificam.

2.2. Métodos de identificação (automática) das relações CST

Há vários trabalhos que propõem métodos de identificação automática das relações da teoria/modelo CST ou de relações semelhantes. Dentre eles, destacam-se: (i) Zhang *et al.* (2002, 2003), Zhang e Radev (2005), MacCartney *et al.* (2006), Miyabe *et al.* (2008) e Kumar *et al.* (2012) para o inglês, e (ii) e Maziero (2012) e Souza (2015), para o português.

Nos métodos de Zhang *et al.* (2002, 2003) e Zhang e Radev (2005), a identificação das relações CST é feita em 2 etapas.

Zhang e Radev (2005) verificam a existência de alguma conexão lexical entre as sentenças que compõem um par. Isso é feito porque é improvável a ocorrência de relações

CST se dê entre sentenças que sejam lexicalmente muito diferentes (ZHANG *et al.*, 2003). Para tanto, aplica-se a medida estatística *word overlap*, que é determinada pela aplicação da fórmula em (4). Caso o valor da *word overlap* obtido seja igual ou superior a 0.12[7], considera-se que as sentenças do par sob análise são relacionadas.

(4)

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

Para calcular a *word overlap* (*Wol*) entre um par de sentenças (S1 e S2) provenientes de textos distintos, mas que tratam do mesmo assunto deve-se, segundo a fórmula em (4), 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 as *stopwords*[8], números e símbolos). O resultado obtido será entre 0 e 0,5, sendo que, quanto mais próximo de 0,5, mais redundantes serão as sentenças do par e, quanto mais próximo de 0, menos redundantes.

Na etapa posterior, o método determina a relação CST que ocorre entre as sentenças do par. Para tanto, observam-se as características (ou atributos) de diferentes níveis linguísticos, a saber: (i) número de palavras idênticas entre as sentenças (atributo lexical), (ii) número de classes de palavras idênticas (atributo sintático)[9], 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[10] (WN.Pr), uma base relacional de dados lexicais (FELLBAUM, 1998).

---

[7] 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 adequado para a detecção da similaridade.

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

[9] 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.

[10] 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.

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

Para avaliar o método, Zhang e Radev (2005) utilizaram um *corpus* de treinamento/teste composto por 6 coleções de textos, cujas principais características estão descritas na Tabela 2[11].

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 *e power lines*	4	21

Tabela 2: *Corpus* de treinamento e teste de Zhang e Radev (2005).
Fonte: Zhang e Radev (2005).

As sentenças das referidas coleções foram manualmente anotadas com relações CST e os atributos necessários para a determinação da similaridade (isto é, *word overlap*) e das relações CST (atributos lexical, sintático e semântico) foram explicitados. 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, mas os referidos atributos foram.

Na sequência, as 7 coleções do *corpus* foram submetidas a algoritmos de AM que, a partir dos atributos explícitos, aprendem padrões estatisticamente relevantes e testados. Isso

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

pode ser feito 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 foram expressos pelas medidas clássicas de avaliação no PLN, a saber: precisão, cobertura e medida-f (HIRSCHMAN; MANI, 2003). No 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.

RELAÇÃO CST	PRECISÃO	COBERTURA	MEDIDA-F
SEM RELAÇÃO	0.8875	0.9605	0.9226
EQUIVALENCE	0.5000	0.3200	0.3902
SUBSUMPTION	0.1000	0.0417	0.0588
FOLLOW-UP	0.4727	0.2889	0.3586
ELABORATION	0.3125	0.1282	0.1818
DESCRIPTION	0.3333	0.1071	0.1622
OVERLAP	0.55263	0.2941	0.3773

Tabela 3: Resultados das medidas de avaliação de Zhang e Radev (2005).
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 ser detectada.

MacCartney *et al.* (2006) identificam automaticamente o relacionamento (*entailment*) estabelecido entre unidades de análise sob a forma de acarretamento. De acordo com os autores, dado um par de sentenças, S1 e S2, o acarretamento ocorre quando uma delas é tida como *hipótese*, enquanto a outra, como *texto*, como exemplificado em (5).

(5)

S1: Estima-se que 2,5 a 3,5 milhões de pessoas morreram de AIDS no ano passado.
S2: Mais de 2 milhões de pessoas morreram de AIDS no ano passado

Em (5), S1 é o *texto*, e S2, a *hipótese*. Por meio de uma implicatura (ou acarretamento) semântica, S2 está contida em S1, já que, de fato, "*mais de 2 milhões de pessoas*" está compreendido por "*2,5 a 3,5 milhões de pessoas*". Assim, essas sentenças estão em relação de *entailment*.

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 por nós e as relações gramaticais estabelecidas entre elas são representadas por arestas. Na sequência, alinham-se os nós correspondentes das sentenças 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 pelo *texto*.

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

a) Polaridade: marcadores linguísticos em contextos de polaridade negativa, expressos pela simples negação (por exemplo, "não"), quantificadores negativos ("menos"), preposições restritivas (como "exceto") e superlativos.
b) Adjunção (do inglês, *adjunct atributes*): marcadores que evidenciam o "abandono" ou a adição de adjuntos sintáticos com características de modificadores, os quais promovem distinções entre sentenças semelhantes (como em "*Os cachorros latem*" que se distingue de "*Os cachorros latem hoje*") ou preservação de sentido (como em "*Os cachorros latem*" tem seu sentido compreendido em "*Os cachorros latem alto*"). Ambas as operações mantém a ligação do texto com a hipótese.
c) Antonímia: marcadores que evidenciam a polaridade (negativa ou positiva) entre um par de antônimos advindos do texto e da hipótese. Para tanto, os autores propõem identificar os antônimos com base na WN.Pr e uma lista de antônimos de referência. Observa-se, dessa maneira, qual polaridade é expressa a partir do par de antônimos por meio do contexto do texto e/ou da hipótese.
d) Modalidade: marcadores que identificam a modalização entre o texto e a hipótese. Os autores analisam seis modalizadores (a saber, *possible, not possible, actual, not actual, necessary* e *not necessary*) e definem cinco 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 (como em "*O ladrão tentou escapar*" que se difere de "*O ladrão escapou*", já que esta sentença pressupõe àquela).

f) Quantificação: marcadores que evidenciam relação de quantificação entre o texto e a hipótese, como em *"Cada empresa deve informar a seus funcionários"* em que se difere quanto ao sentido de *"Uma empresa deve informar a seus funcionários*", já que essa sentença admite a interpretação de apenas uma empresa deve informar seus funcionários, enquanto aquela pressupõe que haja mais de uma empresa). Os autores decompuseram essa categoria em outras cinco classes, a saber, *no*, *some*, *many*, *most* e *all*.

g) Tempo e data: marcadores que evidenciam a relação de tempo/data e entre o texto e a hipótese, como em "*Estima-se que 2,5 a 3,5 milhões de pessoas morreram de AIDS no ano passado.*", em que a expressão destacada evidencia uma informação temporal/de data.

h) Alinhamento: marcadores da qualidade do alinhamento entre o texto e a hipótese, sob os valores de *good score* ou *bad score*. O objetivo dessa observação é representar a qualidade da ligação entre o texto e a hipótese comparando a distância entre os resultados e a referência.

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 (acurácia) e a "pontuação ponderada" como métricas de avaliação dos atributos. Por meio dos atributos levantados, os autores geraram grafos em que cada palavra de uma sentença é mapeada em pares de palavras de outra sentença, ou a nenhuma palavra. A acurácia máxima levantada por MacCartney *et al.* (2006) foi de 0,65, resultado que se mostra positivo à área.

Miyabe *et al.* (2008) também investigaram a identificação automática de relações semelhantes às do modelo CST. Em especial, os autores focaram nas relações *Equivalence* (equivalente a relação *Equivalence* no modelo CST) e *Transition* (equivalente a relação *Contradiction* no modelo CST). Segundo os autores, *Equivalence* ocorre entre 2 sentenças quando estas veiculam a mesma informação por meio de palavras diferentes. *Transition*, por sua vez, ocorre entre 2 sentenças quando estas veiculam a mesma informação, mas apresentam distinção numérica.

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)

Quadro 5: Exemplo de relações semânticas de Miyabe *et al.* (2008)

Fonte: Miyabe *et al.* (2008).

No Quadro 5, de acordo com os autores, há uma relação de *Transition* entre S1 do Texto 1 e S1 do Texto 2 pois, apesar de transmitirem informação similar, apresentam quantidades distintas de usuários por conta da variação de datas em que os eventos ocorrem. Além disso, há uma relação de *Equivalence*, pois as sentenças transmitem a mesma informação por meio de paráfrase.

Para identificar essas relações, os autores consideraram a similaridade entre as sentenças com base em: (i) quantidade de caracteres de cada sentença, (ii) data de publicação do texto-fonte de cada sentença, (iii) posição das sentenças nos texto-fonte, (iv) similaridade lexical (capturada pela medida do *cosseno*[12]), (v) similaridade semântica, (vi) conjunções, (vii) expressões ao final da sentença, (viii) entidade nomeada e (ix) tipo de entidade nomeada ("lugar", "hora", por exemplo).

Miyabe *et al.* (2008) utilizaram um *corpus* que compreende 115 conjuntos de textos jornalísticos inter-relacionados. Os textos foram organizados em 15 coleções, compostos em média por 10 textos cada, em média.

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

Para a identificação da relação *Equivalence*, os autores anotaram um *corpus* utilizando o 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 o seguinte algoritmo[13]: (i) identificar sintagmas nominais (SN) constituídos por valores numéricos, (ii) identificar sentenças em que os valores números são apenas predicativos, (iii) buscar SNs que dependem de predicação e (iv) extrair SNs encontrados em (iii), exceto informações sobre data. No Quadro 9, "*one milion*" e "*1.500,000*" são valores numéricos que predicam o SN ("*number of users*"), e "*had reached*" é a construção verbal que promove a cópula entre eles.

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

Com o objetivo de desenvolver métodos para a identificação automática das relações CST, Maziero (2012) gerou vários classificadores. O melhor deles foi usado para a construção do CSTParser, um analisador discursivo para textos em PB. Nessa ferramenta de PLN, as relações CST são identificadas com base nos atributos linguísticos até então mais difundidos da literatura. Especificamente, o método de análise multidocumento de Maziero (2012) identifica as relações CST com base nas informações descritas no Quadro 6.

DIFERENÇA DE TAMANHO EM PALAVRAS (S1-S2)
PORCENTAGEM DE PALAVRAS EM COMUM EM S1
PORCENTAGEM DE PALAVRAS EM COMUM EM S2
POSIÇÃO DE S1 NO TEXTO (0- INÍCIO, 2- FIM, 1- MEIO)
NÚMERO DE PALAVRAS NA MAIOR *SUBSTRING* ENTRE S1 E S2
DIFERENÇA NO NÚMERO DE SUBSTANTIVOS ENTRE S1 E S2
DIFERENÇA NO NÚMERO DE ADVÉRBIOS ENTRE S1 E S2
DIFERENÇA NO NÚMERO DE ADJETIVOS ENTRE S1 E S2
DIFERENÇA NO NÚMERO DE VERBOS ENTRE S1 E S2
DIFERENÇA NO NÚMERO DE NOMES PRÓPRIOS ENTRE S1 E S2
DIFERENÇA NO NÚMERO DE NUMERAIS ENTRE S1 E S2
SOBREPOSIÇÃO DE SINÔNIMOS ENTRE S1 E S2

Quadro 6: Atributos de Maziero (2012).
Fonte: Maziero (2012).

[13] De maneira geral, algoritmo é uma sequência de passos predeterminados para realizar uma tarefa.

Vale ressaltar que, além das características sentenciais do Quadro 6, 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, isto é, resultantes de diferenças numéricas entre as sentenças de um par. Por exemplo, caso haja um símbolo do tipo hora ("h") (ou medidas como metros, quilômetros, etc.) nas sentenças de um par, verifica-se se os valores vinculados a esses símbolos são iguais ou diferentes. Se diferentes, a regra indica que há uma contradição entre as sentenças.

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

Com vistas à avaliação do *parser*, o *corpus* foi manualmente anotado com relações CST. Para tanto, dividiu-se o CSTNews em duas parcelas (treinamento e aplicação) e equacionou-se a tarefa em duas fases. Na fase de treinamento, os anotadores (linguistas-computacionais) realizaram encontros presenciais e, ao final de cada encontro, as relações CST obtidas por cada anotador eram discutidas pelo grupo para verificar a concordância entre eles. Na etapa de aplicação, o restante do *corpus* foi dividido em três, o que levou os anotadores também a se organizarem em três grupos para que cada um realizasse a anotação de uma das parcelas restantes.

Maziero (2012) obteve a precisão geral de 68,13% para o melhor algoritmo desenvolvido. Essa precisão geral é a média ponderada da precisão dos métodos do Quadro 10 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*[14]. Segundo o autor, essa precisão é considerada boa devido à subjetividade inerente à tarefa de identificação das relações multidocumento.

[14] 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.

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*[15] (juntamente com *Refinament*) foi fundida à relação *Elaboration*.

O método de Kumar *et al.* (2012) pauta-se em 4 características sentenciais: (i) similaridade lexical, capturada pelas medidas distintas *cosseno* e *word overlap*, (i) tamanho das sentenças, (iii) similaridade de sintagma nominal e (iv) similaridade de sintagma verbal. Para avaliar o método, os autores utilizam 476 pares de sentenças para treinamento e 206 pares para teste, todos provenientes do CSTBank[16] (RADEV *et al.*, 2004). O conjunto de teste inclui 100 pares compostos por sentenças sem anotação de relações CST.

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

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

---

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

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

Com o objetivo de descrever especificamente as características do comportamento linguístico da complementaridade, Souza (2015) desenvolveu métodos para a identificação automática dos tipos de complementaridade (temporal e atemporal) e das relações CST que a codificam. Para tanto, o autor descreveu manualmente pares de sentenças anotadas com as relações CST e levantou um conjunto de atributos traduzidos em métodos de identificação, a saber: (i) distância entre as sentenças, (ii) sobreposição de nome, (iii) ocorrência de advérbios temporais em S1, (iv) ocorrência de advérbios temporais em S2, (v) ocorrência de Expressões Temporais em S1, (vi) ocorrência de Expressões Temporais em S2, (vii) sobreposição de subtópicos, (viii) ocorrência de marcador discursivo em S1 e (ix) ocorrência de marcador discursivo em S2.

Souza (2015) utilizou os algoritmos PART (WITTERN; FRANK, 1998), J48 (QUILAN, 1993) e OneR (HOLTE, 1993) para construir classificadores baseados nos referidos atributos da complementaridade.

Na Tabela 4 apresentam-se os resultados das medidas de avaliação (P, C e MF) para cada um dos classificadores na tarefa de identificação das relações CST de complementaridade. Já a Tabela 5 apresentam os resultados dos classificadores na tarefa de identificação dos tipos de complementaridade.

Relação CST	**One-R**			**PART**			**J48**		
	Precisão	Cobertura	Medida-f	Precisão	Cobertura	Medida-f	Precisão	Cobertura	Medida-f
Elaboration	0.58	0.82	0.68	0.59	0.68	0.63	0.6	0.76	0.67
Follow-up	0.73	0.43	0.54	0.71	0.6	0.65	0.7	0.5	0.58
Historical background	0.76	0.72	0.74	0.75	0.75	0.75	0.78	0.75	0.75

Tabela 4: Resultado das medidas de avaliação para identificação das relações de complementaridade.
Fonte: Souza (2015).

Tipo de Complementaridade	**PART**			**J48**			**OneR**		
	Precisão	Cobertura	Medida-f	Precisão	Cobertura	Medida-f	Precisão	Cobertura	Medida-f
Temporal	0.76	0.73	0.74	0.8	0.68	0.74	0.82	0.6	0.69
Atemporal	0.62	0.66	0.65	0.61	0.75	0.68	0.58	0.8	0.67

Tabela 5: Resultado das medidas de avaliação para identificação dos tipos de complementaridade
Fonte: Souza (2015).

Em relação aos demais trabalhos da literatura, o estudo realizado por Souza (2015) destaca-se por ter descrito a complementaridade, resultando em um conjunto de atributos específicos ao fenômeno linguístico. Tais atributos baseiam-se apenas em informações linguísticas superficiais, ou seja, categorias/informações linguístico-estruturais quantificáveis e que emergem na superfície textual. Baseando-se nessa natureza de atributos, obteve, por exemplo, relativo sucesso na descrição da relação *Historical Background*, de acordo com a MF dos classificadores desenvolvidos, já que se trata de uma relação semântica que pode ser capturada por esse tipo de atributos. Entretanto, atributos superficiais não conseguiram caracterizar com o mesmo sucesso as relações *Follow-up* e *Elaboration*, o que pode ser um indicativo de que atributos dessa natureza não são capazes de caracterizar e, consequentemente, identificar completamente tais relações entre sentenças de um par, tendo como resultado de MF de todos os classificadores abaixo dos resultados da relação *Historical Background*.

Partindo-se dessas considerações, acredita-se que uma descrição linguística mais refinada, que leve em conta informações/características linguísticas superficiais e profundas, em especial, possa aprimorar e aprofundar o conhecimento sobre a complementaridade estudada por Souza (2015).

Na próxima subseção, serão apresentados conceitos de Tipo e Gênero Textual, os quais são importantes para a delimitação investigativa deste trabalho. Especificamente, fazem-se considerações tipo textual “informativo”, com foco no gênero textual “notícia”, já que tal tipo e gênero textuais configuram os textos que compõe o *corpus* utilizado nesta pesquisa.

2.3. Identificação de sinalizadores de relações semânticas

Como demonstrado na subseção anterior, os métodos de identificação (automática) de relações CST, em sua maioria, se baseiam no levantamento de conjuntos de sinalizadores (ou atributos) que possam caracterizar tais relações. Tal trabalho já vem sendo estudado em outros modelos de análise discursiva, como a RST. Um desses trabalhos de grande expressividade e destaque é o de Taboada e Das (2013), o qual é resultado de outros trabalhos investigativos anteriores (p.ex. Taboada 2004).

Com o objetivo de levantar atributos que caracterizem as relações RST, os autores partem do pressuposto de que todas as relações desse modelo são sempre marcadas. Isso quer

dizer que sempre haverá evidências perceptíveis (mas nem sempre textuais) a intenção (discursiva e semântica) do autor em sua produção discursiva. Em um texto qualquer, ao utilizar *porém* como conjunção adversativa entre dois períodos, por exemplo, o autor do texto deseja evidenciar um relacionamento de *contraste* entre as duas unidades discursivas; assim, o pesquisador de relações RST irá identificar *porém* como um sinalizador de tal relação.

Desse ponto de vista, os trabalhos que se dedicavam a identificar relacionamentos entre as unidades discursivas no modelo RST denominavam o conjunto pistas textuais como *marcadores discursivos*. Tal nomenclatura faz jus ao que se sabia até a publicação do trabalho de Taboada e Das (2003), já que se baseavam em marcas explícitas no discurso (como conjunções ou certos grupos nominais). Entretanto, ao observarem que tal nomenclatura, de certa maneira, limitava a caracterização de certos relacionamentos semânticos, os autores propuseram apenas *sinalizadores*, pois investigaram atributos que iam além de marcas explícitas ou usuais, até então, no estudo da identificação das relações RST.

Além disso, os autores apontam que a sinalização das relações discursivas deve ser examinada a partir do processamento. Isso quer dizer que Taboada e Das (2013) assumem que as relações semânticas são entidades cognitivas e, então, deve-se descobrir a possível maneira como os interlocutores dos textos (produtores e receptores) são capazes de identificar unidades discursivas com base em pistas linguísticas. Os autores ainda salientam que a comunicação bem-sucedida se baseia em uma interpretação que, relativamente, não dê margem aos equívocos das relações, para as quais são necessários sinais claros. Dessa maneira, a proposta do trabalho em questão é utilizar um *corpus* já anotado com as relações RST com o objetivo de identificar as pistas dessas relações, adicionando informações de como as relações são sinalizadas e acrescentando outros possíveis sinalizadores.

Como resultado, os autores apontaram sinalizadores já previstos em trabalhos investigativos anteriores e acrescentaram uma série deles que não estavam previstos em outras pesquisas, propondo uma tipologia de sinais, reproduzida na Figura 3.

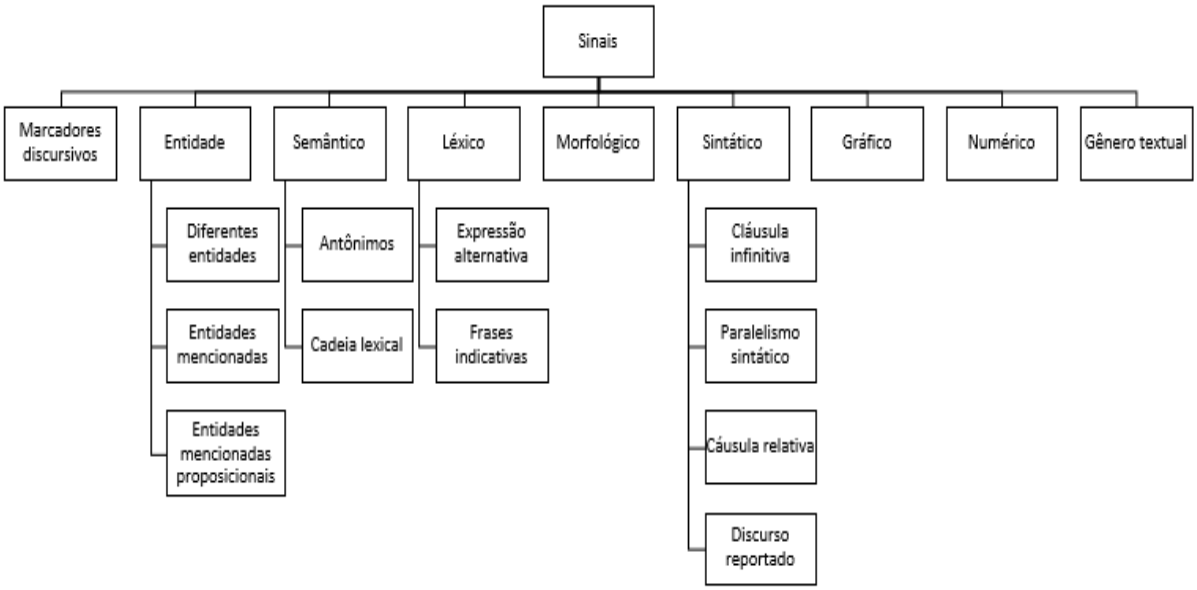

Figura 3: Tipologia de sinalizadores de Taboada e Das (2013).
Fonte: Elaborado pelo autor com base em Taboada e Das (2013).

De acordo com os autores, os sinalizadores se organizam em nove categorias que superordenam outras subcategorias, a saber:

a) *Marcadores discursivos*: sinalizadores que se caracterizam por serem marcas específicas de cada uma das relações RST; em geral, são tidas como expressões léxicas ou conjunções;

b) *Entidade*: sinalizadores que se caracterizam por estabelecem similaridade ou dissimilaridade entre entidades nomeadas de unidades discursivas;

c) *Semântico*: sinalizadores que manifestam relações lexicais (hiperonímia, por exemplo) entre duas entidades de unidades discursivas distintas;

d) *Léxico*: sinalizadores que são traduzidos em palavras que indicam algum tipo de relação, como *acrescentar* que possivelmente indica uma relação *Elaboration*;

e) *Morfológico*: sinalizadores que auxiliam a identificar fatores temporais por meio de desinências verbais;

f) *Sintático*: sinalizadores que indicam relações RST por meio de construções sintáticas específicas, como o discurso reportado;

g) *Gráfico*: sinalizadores que podem indicar relacionamento semântico por meio de pontuações;

h) *Numérico*: sinalizadores que se caracterizam por sinalizar alguma relação RST por meio de especificações numéricas entre unidades discursivas (p.ex. "*João, Pedro e Paulo foram acompanhar Maria no aeroporto*" e "*A garota estava acompanhada de seus três amigos no aeroporto*");

i) *Gênero (textual)*: sinalizadores que evidenciam marcas textuais específicas de cada gênero, como o informativo, em que as primeiras sentenças de um texto desse gênero terão informações genéricas, as quais serão elaboradas/detalhadas nas sentenças subsequentes.

Com base nesse estudo, os autores chegaram a algumas conclusões: (i) há sinalizadores que ocorrem mais em algumas relações RST que em outras (como é o caso de *gênero textual* que ocorre mais na relação *Elaboration*); (ii) há sinalizadores que caracterizam relações RST apenas em combinações com outros (como é o caso de *pontuação* que ocorre juntamente com *sintático* para caracterizar a relação *Background*); (iii), por fim, há sinalizadores que, até então, não tinham sido explorados (como é o caso de *pontuação*).

Salienta-se que as contribuições de Taboada e Das (2013), ainda que para o estudo sistemático das relações RST, foram de suma importância para o levantamento de uma tipologia de sinais que marcassem as relações do modelo semântico CST (tópico que é assunto do Capítulo 4), já que o presente trabalho se filia ao estudo de evidências linguísticas que caracterizam relações semânticas.

Na próxima subseção, apresentam-se algumas considerações sobre tipo e gênero textuais.

2.4. Considerações sobre tipo e gênero textuais

Tendo em vista a investigação de sinalizadores linguístico-estruturais da complementaridade em textos informativos do gênero jornalístico, investigaram-se as características de tal gênero.

Marcuschi (2002), ao construir as definições de tipo e gênero textuais (GT), argumenta que é impossível que haja comunicação verbal (nas modalidades escrita ou oral) que não seja por meio de algum gênero textual. Assim, o autor parte da ideia de que a comunicação verbal se dá somente por algum gênero textual.

De acordo com o autor, *tipo textual* seria uma espécie de construção teórica que é definida pela natureza linguística de sua composição/configuração (p.ex. aspectos lexicais, sintáticos, tempos verbais, etc). Os tipos textuais são categorias fechadas, e limitam-se a tipos narrativos, argumentativos, expositivos, descritivos e injuntivos. Já *GTs* abrangem a noção de características sócio comunicativas que são definidas por conteúdos, propriedades funcionais, estilo e composição próprios. Se por um lado os tipos textuais são categorias limitadas, por outro, os gêneros são quase infinitos, pois representam materialidades (em textos) do comportamento comunicativo da sociedade (p.ex. carta pessoal, romance, bilhete, receitas, notícias, etc).

Mascuschi (2002) constrói um quadro sinóptico, sobre as características definitórias de tipos e gêneros textuais, reproduzido a seguir.

TIPO TEXTUAL	GÊNERO TEXTUAL
Constructos teóricos definidos por propriedades linguísticas intrínsecas	Realizações linguísticas concretas definidas por propriedades sócio comunicativas
Constituição de sequências linguísticas ou sequências de enunciados e não são textos empíricos	Constituição de textos empiricamente realizados cumprindo funções em situações comunicativas
Nomeação que abrange um conjunto limitado de categorias teóricas determinadas por aspectos lexicais, sintáticos, relações lógicas, tempo verbal;	Nomeação que abrange um conjunto aberto e praticamente ilimitado de designações concretas determinadas pelo canal, estilo, conteúdo, composição e função
Designações teóricas dos tipos: narração, argumentação, descrição, injunção e exposição	Exemplos de gêneros: telefonema, sermão, carta comercial, carta pessoal, romance, bilhete, aula expositiva, etc

Quadro 7: Características definitórias de tipo e gênero textuais**.**
Fonte: Marcuschi (2002).

Com base nas definições do Quadro 7, Mascuschi (2002) conclui que o gênero caracteriza-se por ser um evento textual, que tem por definição ser altamente maleável, dinâmico e plástico. Ainda de acordo com o autor, os gêneros podem ser definidos pelos *usos da linguagem* (p.ex. como um falante de um sistema utiliza uma carta de recomendação), pelo *suporte* que veicula o gênero ou por um conjunto de *características formais* (p.ex. traços que diferenciam uma carta de recomendação de uma carta pessoal). Assim, salienta-se que, embora os GTs não possam ser caracterizados nem definidos apenas por aspectos formais, sejam eles estruturais

ou linguísticos, mas sim por aspectos sócio-comunicativos e funcionais, isso não quer dizer que a *forma* seja irrelevante para definir os gêneros.

Ao estudar a estrutura das notícias (em PB), Lage (1993) propôs diretrizes do comportamento linguístico dos textos jornalísticos. Segundo o autor, uma produção textual do gênero *notícia* pode ser organizada com base no método da *pirâmide invertida* (Figura 4).

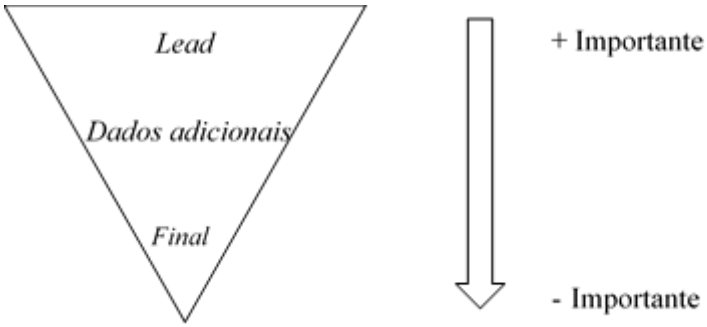

Figura 4: Estrutura do texto jornalístico – Pirâmide invertida.
Fonte: Adaptado de Lage (1993).

De acordo com Lage (1993), o texto de uma notícia é caracterizado por ordenar os fatos de maior a menor importância de um evento. Assim, o autor propõe que a notícia se organiza analogamente a uma pirâmide invertida, em que a informação é veiculada em função da relevância do conteúdo narrado.

A organização proposta, então, se dá em três seções, a saber (i) *lead*, (ii) documentação e (iii) encerramento. O *lead* é o evento principal que está sendo exposto e textualmente corresponde ao primeiro e/ou ao segundo parágrafos do texto, nos quais aspectos informacionais são evidenciados (p.ex. *quem, assunto, quando, onde* e *como* fez/ocorreu o evento exposto). A *documentação* é a complementação do *lead* por meio de detalhamentos das informações já sabidas no início do texto e, textualmente, essa seção ocupa um ou dois parágrafos subsequentes ao *lead*. Por fim, o *encerramento* é a finalização do texto (LAGE, 1993).

Com base em Lage (1993), compreende-se, por exemplo, o motivo de a relação *Historical background* ter baixa frequência no *corpus*, pois o conteúdo capturado por ela não é obrigatoriamente uma parte integrante da notícia. Além disso, informações que precedem o evento principal são mais relevantes na maioria das notícias de caráter diário. Assim,

privilegiam-se informações que manifestam a continuação e/ou previsão do evento que está sendo veiculado, como é o caso do conteúdo rotulado pelas relações *Follow-up* e *Elaboration*.

2.5. Lições aprendidas

Da revisão sobre os trabalhos em que foram propostos métodos automáticos de identificação das relações multidocumento CST ou semelhantes, observa-se que os métodos se pautam 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 níveis. Com relação à complementaridade, sua identificação pauta-se na captura de informações temporais entre sentenças de um par.

Capítulo 3

SELEÇÃO E ESTUDO DO *CORPUS*

Neste capítulo apresento o corpus *jornalístico multidocumento CSTNews, utilizado nesta pesquisa, além de salientar a distribuição do fenômeno o da complementaridade nele. Ademais, apresento análises linguísticas, com base no* corpus*, sobre o fenômeno em função das relações CST que o traduz.*

3.1. Seleção do *corpus*

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

O CSTNews está estruturado em 50 conjuntos (*clusters*) de textos jornalísticos, sendo que cada *cluster* contém notícias que tratam de um mesmo assunto; cada notícia é proveniente de uma fonte jornalística distinta. 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.*

Os *clusters* no CSTNews estão organizados em categorias, cujos rótulos indicam a seção do jornal da qual os textos que os constituem foram compilados. Assim, têm-se as categorias “mundo”, “política”, “cotidiano”, “ciência” e “esporte”. Além disso, cada *cluster* do *corpus*, além de estar ancorado em uma categoria, possui assuntos (ou tópicos de notícia) distintos, como demostrado no Quadro 8.

CLUSTER *(C)*	CATEGORIA	ASSUNTO
C1	Mundo	Acidente aéreo em Bukavu
C2	Política	Eleições presidenciais no Brasil
C3	Cotidiano	Acidente com o avião da TAM
C4	Cotidiano	Alagamentos e chuvas em São Paulo
C5	Cotidiano	Escolha da nova diretoria da ANAC
C6	Cotidiano	O desejo de tornar o Brasil em um "canteiro de obras"
C7	Ciência	Descoberta de um novo planeta
C8	Esportes	Liga Mundial de Vôlei
C9	Política	Operação Dominó em Rondônia
C10	Mundo	Confronto armado entre Israel e o Hezbollah
C11	Cotidiano	Ataques do PCC em São Paulo
C12	Mundo	Ataques em Muttur resultam na morte de volutários de ONG
C13	Mundo	Ataques em Muttur resultam na morte de volutários de ONG[17]
C14	Mundo	Acidente de trem no Cairo
C15	Mundo	Explosão em um mercado de Moscou
C16	Política	Investigação do "esquema sanguessuga"
C17	Política	Eleições presidenciais no Brasil
C18	Mundo	Tiroteio na Universidade de Virgínia
C19	Esportes	Quadro clínico de Maradona
C20	Política	Votação no Senado pelo aumento de CPMF
C21	Cotidiano	Reforma nas pistas do aeroporto de Cumbica
C22	Cotidiano	Deslizamento de terra no Congo
C23	Mundo	Chuvas e enchentes no Reino Unido
C24	Política	Vitória de Fabiana Murer
C25	Esportes	Desempenho do Brasil na Copa América
C26	Mundo	Desastre no México causado por um furacão
C27	Esportes	Goleada do Brasil sobre o Equador
C28	Esportes	Heptacampeonato da seleção brasileira de vôlei
C29	Mundo	Igreja católica paga idenizações por abusos sexuais
C30	Dinheiro	Lucro do Banco Itaú em 2007
C31	Esportes	Vida de Jade Barbosa
C32	Mundo	Falha nuclear em Tóquio
C33	Cotidiano	Lula na abertura da Conferência da ONU
C34	Cotidiano	Fiscalização da Receita Federal
C35	Mundo	Prisão de Juan Carlos R. Abadias
C36	Cotidiano	Morte de Antônio Carlos Magalhães
C37	Cotidiano	Tropa de choque no Maranhão
C38	Esportes	Resultado da natação no Pan-2007
C39	Cotidiano	Acidente aéreo em Congonhas
C40	Política	Posicionamento sobre Renan Calheiros

[17] Os Clusters 12 e 13 retratam o mesmo assunto, mas os textos foram coletados em dias diferentes.

C41	Esportes	Record pan-americano de Thiago Pereira
C42	Política	Relator de Renan Calheiros
C43	Política	Pedido do Conselho de Ética sobre Renan Calheiros
C44	Política	Investigações contra Renan Calheiros
C45	Cotidiano	Roubo do relógio Rolex de Luciano Hulk
C46	Mundo	Terremoto no Japão
C47	Mundo	Operação militar entre Turquia e Israel
C48	Esportes	Vaias ao técnico de vôlei Bernardinho
C49	Cotidiano	Vaias ao presidente Lula
C50	Política	Apoio do governo à CPMF

Quadro 8: Relação de Categoria e Assunto no *corpus* CSTNews.
Fonte: Elaborado pelo autor.

Cada *cluster* do *corpus* é composto por textos-fonte (dois ou três), sumários monodocumento e multidocumento de referência (manuais) e automáticos, alinhamento manual das sentenças dos sumários multidocumento às respectivas sentenças dos textos-fonte e uma série de camadas de anotações linguísticas. Dentre elas, estão: (i) relacionamentos semânticos multidocumento via CST; (ii) anotação de expressões temporais dos textos-fonte; (iii) etiquetação morfossintática (ou *tagging*); (iv) anotação dos sentidos dos substantivos e verbos; (v) anotação de aspectos informacionais nos sumários multidocumento (*o quê*, *onde*, *quando*, por exemplo), (vi) anotação automática dos textos-fonte via RST e (vii) anotação manual de subtópicos informativos em cada texto-fonte do *corpus*.

A anotação CST, em especial, foi realizada semiautomaticamente. Para tanto, Aleixo e Pardo (2008a) revisaram o conjunto de relações CST proposto por Zhang *et at.* (2002). Dessa revisão, os autores concluíram que algumas relações eram redundantes ou não ocorriam nas produções textuais do *corpus*, resultando em um conjunto de apenas 14 relações. Após a definição do conjunto de relações, Aleixo e Pardo (2008b) desenvolveram o editor CSTTool, com a finalidade de facilitar o processo de anotação do *corpus*. A ferramenta possibilita os processos de (i) segmentação dos textos-fonte em nível sentencial, (ii) identificação, em pares, das sentenças lexicalmente relacionadas por meio da medida *word overlap*, e (iii) disponibilização do conjunto de relações CST ao anotador (ALEIXO; PARDO, 2008b).

Uma das etapas mais importantes do processo de anotação de *corpus* é o cálculo da concordância entre os anotadores. Tal etapa verifica se os anotadores estão familiarizados com o fenômeno a ser explicitado, garantindo solidez à tarefa.

Assim, por meio da medida Kappa (CARLETTA, 1996), calculou-se a concordância entre os anotadores do CSTNews, sendo que a anotação das relações *Overlap* e *Elaboration*, que são as mais frequentes no *corpus*, obteve as mais altas taxas de concordância, isto é, 0.562 e 0.321, respectivamente. De acordo com Krippendorff (1980), um resultado abaixo de 0.67 não pode ser considerado confiável. Entretanto, deve-se ter em vista que a anotação CST em questão dispunha de 14 rótulos diferentes, os quais não são mutuamente exclusivos, o que torna a tarefa muito mais complexa, o que justifica as concordâncias obtidas.

Na Tabela 6, apresenta-se o resultado da concordância entre os anotadores e a frequência de cada uma das relações CST no CSTNews. Ressalta-se que algumas relações ocorrem com frequência baixa ou não ocorrem no *corpus* devido a sua tipologia. A relação *Translation*, por exemplo, não ocorre no CTSNews (frequência 0%), posto que se trata de um *corpus* monolíngue.

RELAÇÃO	CONCORDÂNCIA (*Kappa*)	FREQUÊNCIA NO *CORPUS*
Elaboration	0.321	23.98%
Overlap	0.562	19.85%
Subsumption	0.006	15.24%
Background	não informado[18]	6.49%
Atributtion	não informado	5.68%
Equivalence	não informado	5.09%
Follow-up	0.009	4.72%
Contradiction	não informado	4.35%
Summary	0.003	4.35%
Identity	não informado	3.69%
Modality	não informado	3.54%
Indirect Speech	0.013	2.73%
Citation	não informado	0.29%
Translation	não informado	0%

Tabela 6: Relação entre Concordância e Frequência das relações CST no *corpus* CSTNews.
Fonte: Cardoso *et. al.* (2011).

[18] Aleixo e Pardo (2008a) não informam o resultado da medida Kappa para todas relações CST encontradas no *corpus*.

Na próxima subseção apresentam-se os recortes realizados no CSTNews que resultaram na organização de *subcorpora* necessários para o estudo do fenômeno e o desenvolvimento de testes de algoritmos de AM para identificação automática das relações CST de complementaridade.

3.2. Os *subcorpora*

Com o objetivo de estudar especificamente a complementaridade, fez-se um recorte no CSTNews, que consistiu em selecionar apenas os pares de sentenças anotados com as relações CST de complementaridade, ou seja, *Follow-Up, Historical background* e *Elaboration*. Esse recorte foi feito por meio da interface *online* de consulta ao *corpus*[19]. Como resultado, obteve-se um subconjunto de 713 pares de sentenças distribuídos em: (i) 319 pares de complementaridade atemporal (*Elabaration*) e (ii) 313 pares de complementaridade temporal, sendo 260 pares de *Follow-Up* e 73 pares de *Historical Background.*

Na Tabela 7, apresenta-se a distribuição (em pares de sentenças) da complementaridade no *corpus* CST em função da relação entre os *clusters* e as relações CST.

CLUSTER	RELAÇÃO		
	ELABORATION	FOLLOW-UP	HISTORICAL BACKGROUND
C01	11	6	2
C02	9	15	0
C03	11	2	2
C04	6	27	3
C05	1	7	2
C06	17	0	0
C07	14	1	0
C08	9	9	2
C09	14	9	0
C10	6	18	1
C11	7	2	0
C12	5	13	1
C13	7	9	0
C14	2	2	3

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

C15	5	5	0
C16	11	3	0
C17	3	6	0
C18	11	6	6
C19	2	7	0
C20	7	5	0
C21	30	12	0
C22	0	0	0
C23	2	1	2
C24	2	0	1
C25	6	5	1
C26	12	7	0
C27	1	7	0
C28	8	3	6
C29	9	4	14
C30	2	1	0
C31	2	1	0
C32	5	0	7
C33	9	3	0
C34	14	2	0
C35	7	2	8
C36	4	8	1
C37	1	4	1
C38	3	1	4
C39	7	2	0
C40	3	5	0
C41	9	1	1
C42	2	2	0
C43	1	9	0
C44	2	3	0
C45	4	0	0
C46	5	4	4
C47	2	10	4
C48	1	4	0
C49	3	1	0
C50	5	6	0
TOTAL	319	260	73

Tabela 7: Distribuição da complementaridade no *corpus* CSTNews.
Fonte: Elaborado pelo autor.

A partir desse recorte inicial, criaram-se *subcorpora* que atendessem a finalidade desta pesquisa. De acordo com Sardinha (2000), um *corpus* pode ter finalidades de estudo, referência e treinamento (ou teste). O *corpus* de estudo dará subsídio às análises preliminares

do objeto e/ou do fenômeno de estudo. O *corpus* de referência dará suporte a uma análise contrastiva entre este e o conjunto de estudo. O *corpus* de treinamento será submetido a testes que se pautarão no conhecimento levantado a partir das observações realizadas nos *subcorpora* de estudo e referência.

Tendo em vista que no CSTNews há 713 pares de sentenças anotados com as relações de complementaridade, construiu-se um *subcorpus* de estudo em que foi possível observar o comportamento linguístico da complementaridade. Tal como proposto por Sardinha (2000), esse subconjunto permitiu analisar sistematicamente as relações CST de complementaridade e extrair características de cada uma delas. Tais características, posteriormente, transformaram-se em sinalizadores linguístico-estruturais que caracterizaram os pares de sentenças que compuseram esse conjunto. Especificamente, fez-se uma análise manual dos 10 primeiros *clusters* do *corpus*, os quais englobam 204 pares com complementaridade, sendo: (i) 98 pares cujas sentenças estão anotadas com a relação *Elaboration* (isto é, complementaridade atemporal), (ii) 12 pares com a relação *Historical background* e (iii) 94 com a relação *Follow-up* (isto é, complementaridade temporal).

O *subcorpus* de treinamento/teste[20] subsidiou o desenvolvimento de classificadores que se baseiam nos sinais genéricos e específicos advindos da análise manual da complementaridade no *subcorpus* de estudo. Entretanto, há um desbalanceamento entre as relações, o que pode comprometer a qualidade dos classificadores por privilegiar as relações CST que possuem mais pares de sentenças anotados, no caso, *Elaboration* e *Follow-up*.

De acordo com Batista *et al.* (2004), o balanceamento de classes pode ser realizado considerando duas abordagens, a saber:

(i) replicar instâncias das classes menos frequentes até atingir um número próximo à quantidade dos exemplos da classe mais frequente.

(ii) excluir instâncias das classes mais frequentes até atingir um número próximo à quantidade dos exemplos da classe menos frequente.

Tendo em vista que a abordagem definida em (i) pode ocasionar *overfitting* (já que exemplares do conjunto de teste poderiam estar presentes no conjunto de treinamento), optou-se pelo método definido em (ii), baseando-se na relação *Historical Background*, a qual é a

[20] Por conta da técnica de *k-fold cros-validation* (em que *k* representa o número de pastas) em AM (c.f. p.96) para a criação de classificadores, optou-se por não criar o *subcorpus* de referência.

relação que possui menos exemplares anotados. Assim, o *subcorpus* de treinamento/teste foi configurado com 76 pares de sentenças de cada uma das três relações.

Na Tabela 8, apresenta-se a quantificação de cada um dos três *subcorpora* construídos.

RELAÇÃO CST	*SUBCORPUS* DE ESTUDO	*SUBCORPUS* DE TREINAMENTO/TESTE
		QNT. DE PARES
ELABORATION	98	75
FOLLOW-UP	94	75
HISTORICAL BACKGROUND	12	75
TOTAL	204	225

Tabela 8: Dados quantitativos dos *subcorpora*.
Fonte: Elaborado pelo autor.

Em uma fase posterior ao estudo preliminar da complementaridade, analisou-se manualmente todo o *corpus* CSTNews para que a descrição dos sinais linguísticos pudesse ser a mais autêntica e ampla possível. Entretanto, alguns pares de sentença foram excluídos dessa análise, pois a classificação/rotulação das relações CST nesses pares causou dúvidas, como exemplificado em (6).

(6)

S1: De acordo com um porta-voz militar, o Hezbollah contava com pelo menos 11 mil mísseis e foguetes "apontados contra Israel" no começo as hostilidades.

S2: Testemunhas disseram que os disparos de mísseis feitos pelo Hezbollah duraram cerca de 15 minutos.

As sentenças que compõem o par em (6) fazem parte do *Cluster* 10, o qual é composto por textos que noticiam uma explosão em um supermercado moscovita. A primeira sentença relata o poder armamentista do Hezbollah ("*pelo menos 11 mil mísseis e foguetes*"), além de informar que Israel era o alvo. Já a segunda sentença noticia a duração do ataque realizado pelo Hezbollah ("*duraram cerca de 15 minutos*").

Ao reanalisar tais sentenças sob a perspectiva do modelo CST, o par em (6) poderia apresentar uma relação de *Elaboration*, já que S2 acrescenta uma informação de caráter suplementar à S1 (no caso, a duração do ataque). Entretanto, tal par de sentenças, na anotação realizada no CSTNews, foi anotado como *Follow-up*. Para considerar este último rótulo, é necessário partir do princípio que a informação contida em S2 (no caso, o tempo dos disparos feitos pelo Hezbollah) ocorre após o evento narrado em S1, (no caso, o grupo terrorista estar

com mísseis apontados contra Israel "*no começo* das hostilidades). Assim, ter-se-ia a interpretação de que S2 ocorre após S1. Tendo em vista que essa interpretação foi considerada questionável, tal par foi excluído, tal como 24 pares anotados com a relação *Elaboration*, 24 pares com a relação *Follow-up* e 4 pares com a relação *Historical Background*.[21]

De acordo com Aleixo e Pardo (2008a), a concordância da anotação das relações *Elaboration* e *Follow-up* tem Kappa de 0.321 e 0.009, respectivamente. No referido manual, não há o índice Kappa relativo à anotação da relação *Historical background*. Radev *et al.* (2004), ao propor o modelo teórico e realizar a anotação CST em textos em inglês, apresentou 0.4021 de concordância máxima entre os anotadores. Esse índice é considerado baixo no PLN, o que gera a desconfiança quanto à anotação, de acordo com Krippendorff (1980, *apud* ALEIXO; PARDO, 2008a).

Entretanto, deve-se ressaltar que a anotação CST possui alto grau de subjetividade, uma vez que, para cada par de sentenças, o anotador possui vários rótulos possíveis. Além disso, como demonstrado no Quadro 2 (p.22) os rótulos CST não são mutuamente exclusivos, aumentando o grau de complexidade da tarefa, e reforçando que se deve atenuar o fato de a concordância ser baixa.

Assim, optou-se por excluir os pares da análise cujas relações eram duvidosas, já que poderiam comprometer a qualidade da descrição linguística do fenômeno da complementaridade, além da própria geração de classificadores.

Ainda, destaca-se um dos entraves enfrentados durante o processo de estudo e caracterização dos *subcorpora* aqui criados. Da anotação CST que originou o CSTNews, não havia delimitação dos trechos sentenciais que compreendiam a complementaridade, ou que servissem de motivação ao anotador para admitir dado rótulo. Diante disso, identificaram-se manualmente, em todos os pares de sentenças, os respectivos trechos em que ocorriam a informação suplementar. Para tanto, os pares de sentenças receberam delimitadores, como exemplificado em (7).

[21] A distribuição da complementaridade no corpus CSTNews representada na Tabela 7 já não considera os pares de sentenças excluídos e indicados aqui.

(7)

S1: O ministro da Defesa, Nelson Jobim, deve encaminhar o nome da economista Solange Vieira para assumir uma das diretorias da Agência Nacional de Aviação Civil (Anac).

S2: <FU>A Solange vai ser a nova presidente da Anac</FU> - disse Jobim, em jantar que celebrou os 50 anos da Rede RBS em Brasília.

Em (7), o par de sentenças faz parte do *Cluster* 5, cujos textos-fonte informam sobre a indicação de Solange Vieira ao cargo de diretora da ANAC, após um acidente aéreo em São Paulo que levou à morte todos os passageiros e tripulantes da aeronave. No referido par de sentenças, S1 veicula que o ministro da Defesa indicaria Solange Vieira ao cargo de presidente da ANAC, informação que é efetivada pelo conteúdo narrado em S2.

Nesse caso, S2 manifesta a informação complementar, a qual está segmentada pelos delimitadores[22] "<FU>" e "</FU>". Tais etiquetas permitem que a informação complementar seja recuperada de forma automática, caso seja necessário. Ademais, tal delimitação garante que a descrição dos sinais de complementaridade nos pares de sentenças seja referente, de fato, ao segmento em que a informação suplementar e as relações CST ocorrem.

Na próxima subseção apresentam-se análises das características da complementaridade, obtidas com base no *subcorpus* de estudo.

3.3. Análise da complementaridade em *corpus*

De acordo com Maziero *et al.* (2010) compreende-se, de modo geral, a complementaridade pela relação que é estabelecida entre duas sentenças, S1 e S2, sendo cada uma delas provenientes de textos distintos. Além disso, S2 deve apresentar informação complementar em relação a algum elemento presente em S1. Admite-se ainda que as sentenças do par podem compartilhar conteúdo informacional, mas uma delas deve ter alguma informação aditiva que não esteja presente na outra.

De acordo com a tipologia e a definição propostas pelos autores, a complementaridade pode ser determinada pela presença ou ausência de informações temporais. A complementaridade temporal envolve sobreposição de conteúdo entre as sentenças de um par,

[22] Os delimitadores utilizados nos pares de sentenças foram "FU" para *Follow-Up*, "HB" para *Historical Background* e "ELAB" para *Elaboration*.

sendo que S2 apresenta informação adicional ancorada na informação temporal, a qual trata de um acontecimento anterior ou posterior ao evento principal descrito em S1.

Ainda segundo Maziero *et al.* (2010), a complementaridade atemporal também se caracteriza pela sobreposição de conteúdo entre as sentenças de um par, sendo que uma das sentenças fornece informação adicional sobre o tópico principal. No entanto, o que a diferencia da complementaridade temporal é o fato de que a informação adicional não é de natureza temporal, e nem sempre é marcada linguisticamente na superfície textual (MAZIERO *et al.* 2010; SOUZA, 2015; SOUZA; DI-FELIPPO, 2018).

Nas subsseções a seguir, analisa-se o comportamento linguístico-estrutural das complementaridades temporal e atemporal codificadas pelas relações CST com o intuito de levantar os sinais que ocasionalmente são utilizados para delimitar e/ou caracterizar as relações.

3.3.1. Historical Background

De acordo com a definição e a classificação Maziero *et al.* (2010) das relações CST, *Historical background* ocorre quando, em um par de sentenças (S1 e S2), S2 apresenta informações históricas ou passadas sobre algum elemento presente em S1. Os autores ainda apontam que o elemento explorado em S2 deve ser o foco dessa mesma sentença.

Assim, pode-se dizer que há certa sobreposição de conteúdo entre as sentenças do par, sendo que uma delas apresentará uma informação aditiva com foco na eventualidade temporal. Tal sobreposição ocorre por meio do relacionamento dos eventos expressos nas sentenças do par. Isso aponta que tais eventos devem ser distintos entre si.

(8)

S1: A seleção brasileira de vôlei voltou a fazer bonito, desta vez na final da Liga Mundial, disputada contra a Rússia neste domingo no ginásio de Spodekna, em Katowice, na Polônia.

S2: *Sua última derrota em finais da Liga Mundial*, aliás, *ocorreu em 2002*, coincidentemente para a Rússia.

O par de sentenças em (8) faz parte do *Cluster* 28, o qual possui textos sobre a trajetória da seleção brasileira na Liga Mundial de Vôlei. No exemplo, S1 aborda o desempenho da seleção masculina de vôlei na Liga Mundial daquele ano, disputada contra a Rússia, na

Polônia; já S2 informa a derrota do Brasil sobre a Rússia disputando a edição de 2002 do mesmo campeonato.

Assim, o par de sentenças em (8) pode ser anotado com a relação *Historial background*, já que (i) S1 e S2 tratam das mesmas entidades (no caso, as equipes do Brasil e da Rússia) e (ii) S2 apresenta uma informação história sobre a participação do Brasil em uma edição anterior do campeonato.

Além desse tipo de ocorrência, *Historical background* também pode se manifestar textualmente por meio de expressões adjetivas de comparação.

(9)

S1: A Igreja Católica chegou a um acordo financeiro estimado em US$ 660 milhões (aproximadamente R$ 1,2 bilhão) com mais de 500 pessoas que alegam ter sido vítimas de abuso sexual por padres em Los Angeles, nos Estados Unidos.

S2: Este seria o maior pagamento já feito pela Igreja desde que surgiu o escândalo de abuso sexual envolvendo religiosos em 2002 e elevaria o total de indenizações pago pela Igreja desde 1950, nos Estados Unidos, a US$ 2 bilhões (R$ 3,7 bilhões).

O par de sentenças em (9) faz parte do *Cluster* 29, cujo assunto principal é o abuso sexual envolvendo líderes religiosos católicos. Em (9), S1 relata o valor do acordo financeiro estabelecido entre a Igreja Católica e as vítimas do abuso sexual; já S2 informa que, dentre os acordos financeiros já pagos pela Igreja por causa de abusos sexuais, desde 1950, nos Estados Unidos, esse seria o maior.

Assim, o par de sentenças em (9) pode ser anotado com *Historical background*, já que (i) as sentenças relatam de eventos semelhantes (“pagamentos de acordos financeiros sobre abuso sexual de líderes religiosos católicos”), (ii) a ocorrência do pronome “este”, em S2, retoma “acordo financeiro”, em S1, o que revela que ambas as sentenças falam sobre eventos da mesma natureza (no caso, abusos sexuais e indenizações financeiras), (iii) a expressão adjetiva de comparação (“Este seria *o maior pagamento já feito*...”), em S2, reafirma o aspecto relacional entre as duas sentenças e o acréscimo de informação complementar.

Outro fator que é relevante para a caracterização da relação *Historical background* é a distância temporal entre os eventos. Os eventos descritos nos pares de sentenças anotados com tal relação CST não devem ser os mesmos, já que deve haver uma distância temporal entre o conteúdo veiculado pelas sentenças, como exemplificado em (10).

(10)

S1: Naquele horário, segundo a CET (Companhia de Engenharia de Tráfego), havia 110 km de congestionamento em toda a cidade enquanto a média para o horário era de 76km.

S2: Em julho do ano passado, a média foi de 36km no horário.

O par de sentenças em (10) faz parte do *Cluster* 4, o qual trata das consequências das fortes chuvas em São Paulo. No exemplo, S1 informa sobre a extensão do congestionamento ocasionado pela chuva registrada pelo órgão que coordena o trânsito na capital paulista; já S2 relata a extensão média de congestionamento no ano anterior para o mesmo horário.

Assim, o par em (10) foi anotado com a relação CST *Historical background*, uma vez que (i) as sentenças relatam eventos semelhantes ("o congestionamento como consequência das chuvas em São Paulo") e porque (ii) a distância entre os eventos é de um ano, o que reforça o caráter complementar dessa informação temporal.

Por fim, deve-se considerar que, por vezes, as informações temporais nos pares de sentenças anotados com *Historical background* são de aspectos temporais distintos, como demonstrado em (11).

(11)

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.

As sentenças do par (11) fazem parte do *Cluster* 1, cujos textos relatam um acidente aéreo que matou 17 pessoas no Congo. Em S1, há o relato de um acidente aéreo no Congo, sendo confirmado por um porta-voz das Nações Unidas; já em S2, há o histórico de acidentes aéreos na mesma região africana.

Assim, as sentenças representadas em (11) estabelecem a relação *Historical background* entre si porque (i) informam sobre eventos distintos, mas complementares (no caso, *acidentes aéreos* e a *frequência com que acontecem*) e (ii) apresentam aspectos temporais distintos, uma vez que S1 descreve um evento de aspecto *pontual* ("*Um acidente aéreo na localidade de Bukavu*") e S2, uma situação de aspecto *habitual* ("*Acidentes aéreos são frequentes no Congo*").

Na próxima subseção, apresenta-se a análise dos pares anotados com a relação *Follow-up*.

3.3.2. Follow-up

Como apontado por Maziero *et al.* (2010), a relação *Follow-up* ocorre quando, em um par de sentenças (S1 e S2), S2 apresenta acontecimentos que ocorrem após o evento narrado em S1. Os autores ainda ressaltam que os eventos descritos nos pares de sentenças devem ser relacionados e ter ocorrido em um intervalo curto de tempo.

Com base nessa definição, há sobreposição de conteúdo entre as sentenças do par, sendo que uma delas apresenta uma informação aditiva de natureza temporal. Tal sobreposição, como exemplificado em (12), ocorre por meio do relacionamento dos eventos expressos nas sentenças do par.

(12)
S1: Aos 27min, Kaká arrancou e chutou de fora da área.
S2: Kaká acertou um belíssimo chute de longe no ângulo aos 31 e fez 3 a 0.

As sentenças do par exemplificado em (12) fazem parte do *Cluster* 27, o qual engloba notícias sobre a vitória da seleção brasileira de futebol frente à equatoriana. Em S1, relata-se uma jogada feita por Kaká, aos 27 minutos; já S2 informa que o jogador fez um gol logo depois, aos 37 minutos. Assim, ocorre *Follow-up* no referido par de sentenças porque (i) S1 e S2 abordam eventos relacionados à mesma entidade (no caso, "Kaká"); (ii) S2 focaliza fatos que sucedem o evento de S1 e (iii) o intervalo de tempo entre os eventos de S1 e S2 é curto, marcado nas sentenças pelas Expressões Temporais (ETs) (MENEZES-FILHO; PARDO, 2011) "27 minutos" e "aos 31", respectivamente.

Ainda sobre o exemplo (12), percebe-se que os verbos em S2 estão flexionados no *pretérito perfeito* ("acertou" e "fez"), o qual expressa ações já finalizadas. Contudo, a relação *Follow-up* pode se manifestar entre sentenças cujos verbos estão flexionados no *futuro do pretérito*, o qual expressa possibilidades ou previsões de ações, como exemplificado em (13).

(13)
S1: De acordo com a pesquisa, Lula (PT) tem 44% das intenções de voto, contra 25% de Geraldo Alckmin (PSDB) e 11% de Heloísa Helena (PSOL).
S2: O presidente teria 53% das intenções de voto contra 30% de Heloísa.

O par de sentenças em (13) veicula a informação sobre a corrida presidencial no Brasil em 2006, especificamente o resultado das pesquisas de intenções de voto de alguns candidatos. Não há uma ET expressa textualmente que marque a sucessão dos eventos, como em (12). Por outro lado, a construção frástica de S2 é feita por meio do futuro do pretérito ("O presidente *teria*..."), em que se projetam os resultados dos candidatos Lula e Heloísa Helena em um possível segundo turno eleitoral.

Do ponto de vista temporal, os exemplos em (12) e em (13) encaixam-se na definição de *Follow-up*. Entretanto, do ponto de vista da eventualidade, há dois tipos aspectuais distintos que devem ser considerados: em (12), há eventos que de fato aconteceram sucessivamente e em (13), apesar de se narrar eventos subsequentes, o evento em S2 está condicionado, sendo uma possibilidade. De acordo com a análise realizada, percebe-se que tal manifestação de *Follow-up* acontece em *clusters* cujos assuntos principais são política, desastres naturais e esporte, em que informar sobre a possibilidade de um evento é relevante.

Atrelado ao tempo em que os eventos ocorrem, os valores numéricos indicam que há modificação sobre as intenções de voto para os candidatos à presidência ("Lula tem *44%* das intenções de voto, contra [...] *11%* de Heloísa Helena" → "O presidente teria *53%* das intenções de voto contra *30%* de Heloísa"). Tal traço também pode corroborar a ideia de sucessão dos eventos narrados nas sentenças do par em (13).

Além disso, a anáfora associativa ("o presidente", em S2, que retoma "Lula", em S1), as diferentes referências a uma mesma entidade nomeada ("Heloísa Helena" e "Heloísa") e o conhecimento de mundo sobre *política* (construído pelas expressão e palavras "*intenções de voto*" e "*pesquisa*", respectivamente) auxiliam na construção do relacionamento de sentido entre as sentenças do par.

Ainda de acordo com a análise realizada, constatou-se que *Follow-up* também se manifesta em pares de sentenças que apresentam acontecimentos resultantes de um evento principal não presente nas sentenças, como exemplificado em (14).

(14)

S1: O pico de lentidão foi registrado às 9h, com 113 km de lentidão, o dobro do registrado neste horário.

S2: Não havia registro de acidentes graves, ainda às 9h30.

O par de sentenças em (14) faz parte do *Cluster* 4, cujo assunto principal é a consequência das chuvas na cidade de São Paulo. Enquanto S1 narra o congestionamento na cidade, S2 aborda a ausência de acidentes meia hora após o evento de S1 (informação expressa pelas ETs "*às 9h*", em S1, e "*às 9h30*", em S2).

Os eventos narrados em ambas as sentenças são distintos, mas são compreendidos como subeventos ligados à forte chuva que caiu na capital paulista, que seria uma espécie de "supraevento". Assim, compreendendo que os eventos narrados em S1 e em S2 são consequências da chuva, ocorre *Follow-up* nesse par de sentenças porque o evento de S2 acontece após o evento de S1.

Além disso, somente após a leitura dos respectivos textos-fonte das sentenças, é possível associá-las ao mesmo evento (no caso, "chuva forte em São Paulo"), já que não há pistas linguísticas na superfície textual das sentenças do par que possa relacioná-las uma à outra, como traços anafóricos: o fato de não haver registro de acidentes graves poderia estar relacionado a outro evento, como o relato de um furacão, por exemplo. Assim, deve-se considerar a *leitura do cluster* como um traço importante para a compreensão da anotação CST em pares com pouco relacionamento lexical entre as sentenças, como em (14).

Observou-se que *Follow-up* também ocorre entre sentenças em um par, em que S2expressa a efetivação de evento(s) previsto(s) em S1, como demonstrado em (15).

(15)

S1: O ministro da Defesa, Nelson Jobim, deve encaminhar o nome da economista Solange Vieira para assumir uma das diretorias da Agência Nacional de Aviação Civil (Anac).

S2: O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac).

Em (15), as sentenças do par fazem parte do *Cluster* 5, cujos textos relatam sobre a indicação de Solange Vieira à chefia da Agência Nacional de Aviação Civil (ANAC). S1 relata a possibilidade de Jobim indicar Solange Vieira ao cargo ("*deve encaminhar o nome da economista Solange Vieira...*"); já S2 narra que a economista foi, de fato, indicada ao fato ("*O ministro da Defesa, Nelson Jobim, informou...que Solange Vieira...será a nova presidente...*"). Assim, o evento de S2confirma o evento veiculado por S1.

Por fim, há casos em que a relação *Follow-up* é tida em pares de sentenças em que uma delas apresente-se como consequência de uma causa lexicalizada, como em (16). O fato narrado em S1 (no caso, a explosão e o incêndio da aeronave acidentada) resulta na consequência do evento narrado em S2 (no caso, não haver sobreviventes).

(16)

S1: O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.

S2: "Não houve sobreviventes", disse Okala.

Na próxima subseção, apresenta-se a análise dos pares anotados com a relação *Elaboration*.

3.3.3. Elaboration

De acordo com Maziero *et al.* (2010), a relação de complementaridade atemporal, codificada por *Elaboration*, ocorre quando, em um par de sentenças (S1,S2), S2 tem a função de detalhar, refinar ou mesmo elaborar algum elemento de S1. Além disso, S2 não deve repetir informações que estão presentes em S1. Com relação à sobreposição de conteúdo, a definição proposta por Maziero *et al.* (2010) prevê que essa relação pode ocorrer em grau elevado ou não. O par de sentenças em (17) exemplifica a ocorrência de tal relação CST.

(17)

S1: Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.

S2: O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.

As sentenças do par (17) expõem um acidente aéreo que ocorreu no Congo. S1 veicula o número de pessoas que morreram no evento em questão, além de relatar sobre o acidente aéreo em si. S2 relata o que aconteceu com a aeronave após a queda (no caso, uma explosão seguida de incêndio). Assim, S2 fornece detalhes do que aconteceu com a aeronave após a queda, causando a morte das pessoas. Além disso, o verbo "*acrescentou*" auxilia a identificar S2 como veículo de informação adicional à S1.

A complementaridade sem foco na informação temporal também parece ser expressa por traços sintáticos capazes de detalhar os tópicos que estão sendo narrados, como predicativo do sujeito. Percebe-se tais apontamentos em (18).

(18)
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.

O par de sentenças em (18) também faz parte do *Cluster* 1, cujo eixo temático é o acidente aéreo no Congo. S1 apresenta a quantidade e o perfil das pessoas que morreram no evento em questão. Já S2 acrescenta a nacionalidade dos membros da tripulação que, no caso, eram russos.

A informação complementar em (18) é construída por meio da retomada da expressão "*membros da tripulação*" de S1 que, em S2, passa a ser o *tópico frasal*, ocupando a posição sintática de sujeito. Além disso, a informação sobre a nacionalidade é expressa por um *predicativo do sujeito* (no caso, "de nacionalidade russa").

Observou-se também que há recorrência da característica *topicalização* de informação em sentenças anotadas com a relação *Elaboration*, como exemplificado em (19).

(19)
S1: Heloísa Helena, candidata à Presidência pelo PSOL, aparece em terceiro, com 11% das intenções de voto, seguida por Cristovam Buarque (PDT) e Luciano Bivar (PSL), ambos com 1%.
S2: O senador Cristovam Buarque, candidato pelo PDT, teve alta de 29% para 32%.

Em (19), o assunto principal é a corrida presidencial no Brasil, em 2006. S1 narra detalhes sobre as intenções de voto da população acerca dos candidatos Heloísa Helena, Cristovam Buarque e Luciano Bivar, enquanto S2 apresenta a atualização dos dados somente acerca a Buarque.

As sentenças em questão não sobrepõem conteúdo informacional. No entanto, S2 se caracteriza por apresentar informação adicional em relação a S1. Assim, é possível explicar que essa relação de complementaridade é construída pela retomada como tópico, em S2, daquilo que é tido como comentário em S1 (no caso, o candidato Cristovam Buarque), além de acrescentar novas informações sobre o tópico. Tem-se, então, nesse contexto de pares de

sentenças, dois tipos de informações: uma que já é sabida (apresentada em S1) e a outra que é nova ou complementar (apresentada em S2).

Nessa perspectiva, Ilari (1992) define conceitos que explicam as estruturas oracionais em PB, ao apontar que elas se organizam em Tema e Rema dada situação comunicativa (nas modalidades oral ou escrita) entre interlocutores (produtor e receptor/destinatário). Ao tema atribui-se a noção de "informação velha", já que a informação preterida pelo produtor já é de conhecimento do receptor dentro de uma situação comunicativa. Ao rema, atribui-se a noção de "informação nova", já que a informação veiculada não é de conhecimento prévio do receptor.

Assim, em uma análise das relações CST, em que as sentenças que compõem um par são provenientes de textos distintos, aquela que é considerada a primeira sentença necessariamente conterá o Tema, enquanto a segunda, o Rema. A manifestação textual desses conceitos se dá, em geral, por alguma manobra anafórica *direta* (p.ex. pronominal) ou *indireta* (p.ex. associativa).

Por fim, a relação *Elaboration* pode se manifestar em pares de sentenças que tenham focos argumentativos distintos, como exemplificado em (20).

(20)

S1: Nenhuma partida ou chegada internacional, segundo os painéis da Infraero, estavam fora do horário, o que não ocorria com os voos domésticos.

S2: As informações da Infraero não batem com as do painel das companhias aéreas, são 20 partidas atrasadas e 24 pousos atrasados.

O par de sentenças em (20) faz parte do *Cluster* 22, cujos textos-fonte noticiam a reforma nas pistas do aeroporto de Congonhas, em São Paulo. S1 aborda o fato de que não houve atraso nos voos internacionais, de acordo com a Infraero; já S2 aponta que havia atrasos entre partidas e chegadas de voos. Dessa maneira, as sentenças do par argumentam de maneiras distintas (sob a forma de contradição) sobre as informações disponibilizadas pela Infraero.

Na próxima subseção, apresenta-se a caracterização das relações CST de complementaridade em função dos sinais que as marcam.

Capítulo 4

DESCRIÇÃO E PROPOSIÇÃO DA TIPOLOGIA DE DISPOSITIVOS DE SINALIZAÇÃO

Neste capítulo, apresento a tipologia de sinais genéricos e específicos que caracterizaram a complementaridade e, consequentemente, subsidiaram o conjunto de atributos linguístico-estruturais dos classificadores de AM (c.f. Capitulo 5). Para tanto, a partir da descrição dos dispositivos de sinalização (c.f. Capítulo 3, Seção 5), apliquei ao subcorpus *de Teste todos os Sinais que demonstraram relevância quantitativa em pares de sentenças anotados com as relações CST de complementaridade, a saber* Elaboration, Follow-up *e* Historical Background.

4.1. Seleção e delimitação de dispositivos de sinalização da complementaridade

Consoante a Taboada e Das (2013), buscaram-se neste trabalho os *sinais* ou *dispositivos de sinalização* da complementaridade no corpus CSTNews. Entende-se por *sinais*[23] as pistas que indicam e/ou caracterizam a presença de uma das três relações CST de complementaridade.

Essa terminologia é sugerida por Taboada e Das (2013), já que a denominação *marcadores discursivos* é utilizada em trabalhos que visam à identificação de relações RST. Os marcadores delimitam apenas um dos possíveis tipos de sinais que podem ser utilizados para a caracterização de uma relação semântica. Assim, os dispositivos de sinalização

[23] Outros termos sinônimos são *dispositivos de sinalização*, *sinalizadores* ou *marcas*.

englobam *conjunções*, que podem, por exemplo, revelar uma relação causal entre duas proposições, até *vírgulas*, que evidenciam apostos, os quais podem carregar detalhes da informação principal veiculada no par de sentenças.

A análise das características da complementaridade por meio de dispositivos de sinalização ainda permite considerar que tal fenômeno, apesar de ocorrer em textos reais/autênticos, não é apenas construído em nível discursivo, mas também em um nível semântico-pragmático que, por vezes, pode ser materializado textualmente. Dessa maneira, é necessário considerar que analisar relações CST somente quanto à presença ou ausência de marcadores discursivos é limitar o estudo ao que somente se pode capturar textualmente, já que alguns dos sentidos das relações de complementaridade são construídos implicitamente em alguns casos.

Diante disso, cada par de sentenças do *subcorpus* de estudo foi inicialmente analisado com o objetivo de identificar todo e qualquer tipo de sinal que podesse indicar a presença de uma relação de complementaridade nos trechos previamente delimitados. Para tanto, os pares foram dispostos em tabelas semelhantes ao Quadro 9.

PAR	SENTENÇAS	RELAÇÃO	CLUSTER	TRAÇO COMPLEMENTAR	SINAIS GENÉRICOS	SINAIS ESPECÍFICOS
1	S1: O ministro da Saúde egípcio, Hatem El-Gabaly, informou nesta segunda-feira que 57 pessoas morreram e 128 ficaram feridas no choque entre dois trens de passageiros no delta do Nilo, ao norte do Cairo. S2: <HB>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, lotado de passageiros, e que deixou 376 mortos</HB>, segundo números oficiais.	HB	14	Pontual	Temporal	Expressão temporal
					Sintático	Expressão comparativa
						Discurso reportado
					Semântico	Similaridade de evento

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. S2: <HB>Acidentes aéreos são frequentes no Congo</HB>, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.	HB	28	Repetição	Semântico Sintático Temporal	Similaridade de eventos Tema-Rema Expressão temporal

Quadro 9: Exemplo da caracterização do subcorpus de estudo.
Fonte: Elaborado pelo autor.

No Quadro 9, ilustra-se a caracterização dos pares de sentenças do *subcorpus* de estudo. A caracterização está organizada em 7 colunas que representam (1ª) o par, (2ª) as sentenças, (3ª) a relação de complementaridade, (4ª) o *cluster*, (5ª) traço complementar (TC), (6ª) sinal e (7ª) sinal específico.

No Quadro 9, o par 1, por exemplo, possui o TC *pontual*, já que a informação complementar em S2 se refere a um evento que aconteceu em um passado distante daquele que é narrado em S1. Já o par 2, por exemplo, possui o TC *repetição*, uma vez que S2, em relação a S1, veicula a informação sobre eventos que são habituais, por se repetirem com frequência.

Na coluna em que se representa os *sinais genéricos*, procurou-se registrar todos os macroníveis de análise linguística e/ou estrutural nos quais a complementaridade se manifesta. Tais sinais têm caráter genérico, como *pontuação*, *sintaxe* ou *morfologia*. Na coluna *sinais específicos*, registram-se as especificidades de cada um dos sinais genéricos, como *vírgula*, *orações aditivas* ou *advérbio*.

No par 1, do Quadro 9, por exemplo, um dos sinais que captura a complementaridade na relação *Historical background* é *Informação temporal*, sendo identificado especificamente por uma ET (no caso, "*em fevereiro de 2002*"). Além disso, tal complementaridade no

referido par pode ser identificada por um sinal de *Sintaxe*, no caso, uma *Oração comparativa* ("*a maior tragédia ferroviária na história do Egito*").

Na próxima subseção, explica-se cada um dos sinais genéricos e os possíveis sinais específicos que os compõem.

4.2. Descrição dos dispositivos de sinalização da complementaridade

Nesta subseção, descrevem-se como os sinais genéricos e específicos foram efetivamente verificados entre as sentenças de cada par.

A verificação da ocorrência dos dispositivos de sinalização foi realizada manualmente. Com base nos trabalhos de Souza (2015) e Taboada e Das (2013), elencou-se um *hall* de atributos que potencialmente poderiam ocorrer nas relações CST de complementaridade. Esses atributos foram dispostos em planilhas no formato (.xlmx), verificando-se se esses dispositivos ocorriam, de fato, nos pares de sentença anotados com as relações investigadas.

A fim de manter a sistematicidade do estudo e garantir certa objetividade, a anotação foi realizada *cluster* a *cluster*. Além disso, ao identificar, no *subcorpus* de estudo, um novo dispositivo não previsto nos trabalhos apontados, verificavam-se os pares anteriores à identificação a ocorrência do dispositivo recém identificado. Após o término do estudo no referido *subcorpus*, observou-se que alguns dispositivos levantados por Souza (2015) e Taboada e Das (2013) não puderam ser identificados ou delimitados no estudo aqui realizado; porém, quanto a ocorrência dos sinalizadores, observou-se estabilidade, levando a conclusão que o *hall* de dispositivos poderia ser verificado no subcorpus de treinamento/teste.

Escolher metodologicamente por levantar de maneira investigativa um conjunto de dispositivos que potencialmente caracterizam a complementaridade rendeu à anotação do *subcorpus* de estudo mais tempo e atenção ao trabalho (manual e individual) empreendido, totalizando um pouco mais de oito meses. Para realizar a anotação em todos os pares anotados com relações CST foram necessários cerca de mais dez meses de anotação.

Outro ponto também relevante à identificação de dispositivos de sinalização da complementaridade e, consequentemente, a classificação deles em categorias genéricas e específicas pode romper com a ideia de discurso, no PLN. Quanto ao discurso, entende-se um texto, apenas. Dessa maneira, apontar que há uma anáfora entre "o ex-presidente" e "Lula",

por exemplo, é, metodologicamente, coerente. Entretanto, apontar que há anáfora entre segmentos que compõem um par de sentenças que provêm de textos distintos é apontar um relacionamento linguístico interdiscursivo. Apesar disso, optou-se por manter as categorias Anáfora e Sintaxe, por não haver terminologias mais adequadas, até o momento, e por compreender que, no par de sentenças estudado, linguisticamente, uma anáfora nominal pôde ser observada, por exemplo.

Por fim, salienta-se que a descrição dos dispositivos de sinalização da complementaridade foi possível a partir da delimitação do segmento textual em que houvesse informação suplementar, observando como a informação poderia ser capturada no um par de sentenças. Diante disso, verificaram-se as categorias descritas a seguir.

4.2.1. Sinais Anafóricos

As resoluções anafóricas nos pares de sentenças de complementaridade contribuem para a caracterização do fenômeno linguístico estudado já que é possível recuperar o assunto ou referente a que se veicula a informação principal e atribuir-lhe detalhes. Assim, consideraram-se dois tipos de resoluções anafóricas: (i) associativa (p.ex. “Lula” e “ex-frente de grefe”) e (ii) direta (p.ex. “Lula” e “Lula”) (MARCURSCHI, 2000).

(21)

S1: Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.

S2: O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.

O par de sentenças representado em (21) relata um acidente aéreo no Congo. S1 informa que a aeronave se acidentou em meio a uma tempestade durante uma manobra de aterrissagem, em Bukavu; já S2 acrescenta que o avião explodiu e se incendiou após a queda. Em S2, o referente topicalizado “*avião*” é retomado por meio de uma anáfora direta, em que a sentença veicula a informação complementar descrita.

4.2.2. Sinais Estruturais

Observou-se que há casos em que a informação complementar apenas é acessada e/ou construída porque os anotadores se basearam na leitura dos textos-fonte, como exemplificado em (22).

(22)

S1: Na madrugada, a temperatura foi de 12 graus nos lugares mais frios da cidade.

S2: O CGE (Centro de Gerenciamento de Emergências) da Prefeitura de São Paulo registrava oito pontos de alagamento na cidade, às 9h30 desta segunda-feira.

As sentenças em (22) fazem parte do *Cluster* 4 do CSTNews, e noticiam as consequências de uma chuva torrencial na cidade de São Paulo. S1 narra uma das consequências da chuva, que é a baixa temperatura; já S2 narra sobre pontos de alagamento na cidade.

O par em questão foi anotado com a relação *Follow-up*, pois o fato narrado em S2 sucede o evento de S1, já que em S2 tem-se as ETs "*às 9h30*" e, em S2, "*na madrugada*". Entretanto, na superfície textual, não há evidências que indicam essa complementaridade, já que o fato de haver baixa temperatura não está necessariamente ligado ao fato de ocorrer chuva ou alagamento. Portanto, somente com a leitura prévia dos textos-fonte do *cluster* aponta-se que há complementaridade do tipo *Follow-up*.

Assim, em pares de sentenças em que não foi possível determinar sinais linguísticos para a caracterização da informação complementar, considerou-se a *Leitura do cluster* como um sinal que apontaria tal fenômeno.

4.2.3. Sinais de Pontuação

Os sinais de pontuação foram considerados na observação do fenômeno da complementaridade porque reforçam ou destacam informações que detalham o evento narrado como principal. Salienta-se que, nesse caso, somente a ocorrência de sinais de pontuação, como vírgula e parênteses, não é determinante para a caracterização do fenômeno em si ou de uma relação de complementaridade. Antes, é necessário observar a ocorrência desse tipo de informação atrelada a outra informação linguística, como a presença de um aposto, que, no caso, é de natureza sintática.

(23)

S1: Os mesmos parlamentares fizeram, também, um conluio com o Ministério Público e com a Justiça do Estado de Rondônia

S2: Entre os presos estão o presidente do TJ (Tribunal de Justiça) de Rondônia, desembargador Sebastião Teixeira Chaves, e o presidente da Assembleia Legislativa, deputado José Carlos de Oliveira.

Em (23), tem-se um par de sentenças anotado com a relação *Follow-up*, em que se narra uma investigação da Polícia Federal, em Rondônia. Em S2, as vírgulas reforçam a presença de informações aditivas sob a forma de aposto (no caso, "*desembargador Sebastião Teixeira Chaves*" e "*deputado José Carlos de Oliveira*") em relação a S1.

Dessa maneira, observou-se que, no texto jornalístico, *aspas, parênteses e vírgulas* podem marcar informações extras ou detalhes de um evento (como apostos ou adjuntos), bem como as *aspas* podem marcar a presença do discurso direto, evidenciando-se como características da complementaridade. Entretanto, apesar de compreender sua relevância linguística como característica do fenômeno em questão, estatisticamente as pontuações não são relevantes, dada a baixa ocorrência no *corpus* CSTNews. Como resultado, tais sinais não compõem a tipologia de dispositivos que será utilizada para elaborar os classificadores em AM.

4.2.4. Sinais Morfológicos

Observou-se que uma das camadas de análise linguística que pode ser considerada como dispositivo de sinalização da complementaridade diz respeito à morfologia. Tal ocorrência se dá por meio classes de palavras específicas, as quais sinalizam (i) acréscimo ou atualização informativa sobre o mesmo evento (como os numerais que apontam novos resultados da corrida presidencial) ou (ii) informações temporais (como é o caso das desinências temporais dos verbos).

Salienta-se que, mesmo alguns sinalizadores da classe morfológica ocorrendo no *corpus* CSTNews, dada a baixa ocorrência, não foram considerados para compor a tipologia, como é o caso de adjetivos (4 ocorrências) e pronomes (10 ocorrências).

4.2.5. Sinais Sintáticos

Há ocorrências da complementaridade que podem ser capturadas por meio de sinalizadores sintáticos. De acordo com a tradição gramatical, há sinalizadores sintáticos que já possuem a função de acréscimo de detalhes a um evento que esteja sendo narrado. Assim, no período simples, captura-se por meio de adjuntos adverbiais, por exemplo, e no período composto por meio de orações aditivas, por exemplo. Além disso, consideraram-se deslocamentos sintáticos de tópico sentenciais, capturado por meio do deslocamento Tema-Rema.

(24)

S1: Lula disse que a prioridade é a realização de obras nas regiões metropolitanas de grandes centros urbanos.

S2: Lula ressaltou que algumas das obras a serem anunciadas já estão em andamento e outras devem começar de imediato.

Em (24), há um par de sentenças anotado com a relação *Elaboration*. Nele, narra-se a fala do presidente Lula, o qual objetivava revitalizar a infraestrutura do Brasil, transformando o país em um "canteiro de obras". S1 veicula a informação das obras nas regiões metropolitanas do Brasil; enquanto S2 expõe o andamento de algumas obras e o início de outras. O segmento textual "*e outras devem começar de imediato*", em S2, é classificado sintaticamente como uma oração coordenada aditiva, a qual complementa a informação obtida a partir do segmento anterior ("*Lula ressaltou que algumas das obras a serem anunciadas já estão em andamento*") e a informação obtida a partir de S1.

Devido à baixa ocorrência, eliminou-se da tipologia os dispositivos com ocorrência abaixo de 15, a saber: adjunto adnominal (10 ocorrências), aposto (14), orações adverbiais (13), adversativa (6), alternativa (1), apositiva (1), causativa (3), comparativa (2), concessiva (6), conclusiva (1) e predicativa (1).

4.2.6. Sinais Semânticos

Observou-se que a complementaridade também pode ser capturada por meio de sinalizadores de nível semântico. Tais como algumas informações sintáticas, alguns traços semânticos revelam o caráter de adição ou de detalhe ao evento tido como principal. Semanticamente, essas informações podem ser capturadas por meio de relações de sentido, como a *hiponímia*.

(25)

S1: 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.

S2: A falha no reversor --mecanismo que ajuda o avião a frear-- foi detectada pelo sistema eletrônico de checagem da própria aeronave, que continuou voando nos dias seguintes, com o reversor direito desligado.

O par de sentenças em (25) evidencia uma relação *Elaboration*, fornecendo informações sobre um acidente com uma aeronave da TAM, em São Paulo. S1 descreve que um dos

componentes do avião necessário para o pouso (no caso, *o reverso*) estava desligado, enquanto que S2 aponta que a falha nesse componente foi detectada automaticamente. Assim, S2 retoma o tópico de S1 (*Tema*), atribuindo-lhe detalhes (*Rema*)

Do ponto de vista da organização tipológica, o nível semântico, nesta análise, foi subdividido em (i) campo semântico, (ii) semântica lexical (que compreende as relações causa-efeito, parte-todo e hiponímia), (iii) sentido de acréscimo (que compreende a semântica local) e (iv) temporal. Em especial a informação temporal foi alocada à semântica por entender-se que as Expressões Temporais (ET)[24] (MENEZES FILHO; PARDO, 2011), na verdade, são entidades mencionadas de tempo, assim, uma categoria semântica.

Observou-se que, apesar de ocorrerem em pares de sentenças com complementaridade, há sinais que tiveram baixa frequência, a saber: relações semânticas de generalização (7 ocorrências) e sinonímia (12), e papel temático (7). Tais sinais, devido à baixa relevância quantitativa, não foram utilizados na geração dos classificadores de AM.

4.2.7. Sinais Pragmáticos

Na anotação do CSTNews, algumas relações de Zhang *et al.* (2002) foram suprimidas. Assim, pode-se dizer que a definição das relações CST de complementaridade, proposta por Maziero *et al.* (2010), é genérica e não evidencia muitas das especificidades referentes a manifestações linguísticas e estruturais desse fenômeno. Dado isso, propôs-se analisar os traços complementares (TC), compreendidos no nível pragmático de análise linguística.

Os TCs permitem observar possíveis nuances de interpretação da complementaridade, com o objetivo de observar especificidades que caracterizam o fenômeno em cada uma das relações CST que o codifica. Nas relações *Historical background* e *Follow-up*, o TC permite observar o tempo e a duração do evento narrado (p.ex. se *Pontual* ou *Frequente*). Na relação *Elaboration*, o mesmo sinalizador permite observar a estruturação da informação complementar nas sentenças do par por meio, por exemplo, da estrutura *Tema* e *Rema*.

[24] Baseados na proposta de classificação de ETs de Hagège *et al.* (2008, 2010), Menezes Filho e Pardo (2007) realizam a identificação e classificação das ETs no corpus CSTNews. Entretanto, notou-se que alguns advérbios de tempo simples (como "amanhã"), ou nomes de eventos esportivos como tal (como "Jogos Olímpicos de 2007) não estavam cobertos pela anotação. Assim, considerou-se como ET todas as possibilidades previstas em Hagège *et al.* (2008, 2010), ainda que não anotadas no *corpus* CSTNews.

4.3. Proposição da tipologia de dispositivos de sinalização da complementaridade

Consonante ao trabalho desenvolvido por Taboada e Das (2013), que focou na análise de textos anotados com o conjunto de relações RST e sinalizadores evidenciavam que tais relações, apresenta-se aqui uma proposta de organização tipológica dos dispositivos de sinalização da complementaridade. Assim, a Figura 5 ilustra a proposta de tipologia dos dispositivos de sinalização da complementaridade com base nas indicações da literatura e nas descrições do *subcorpus* de estudo.

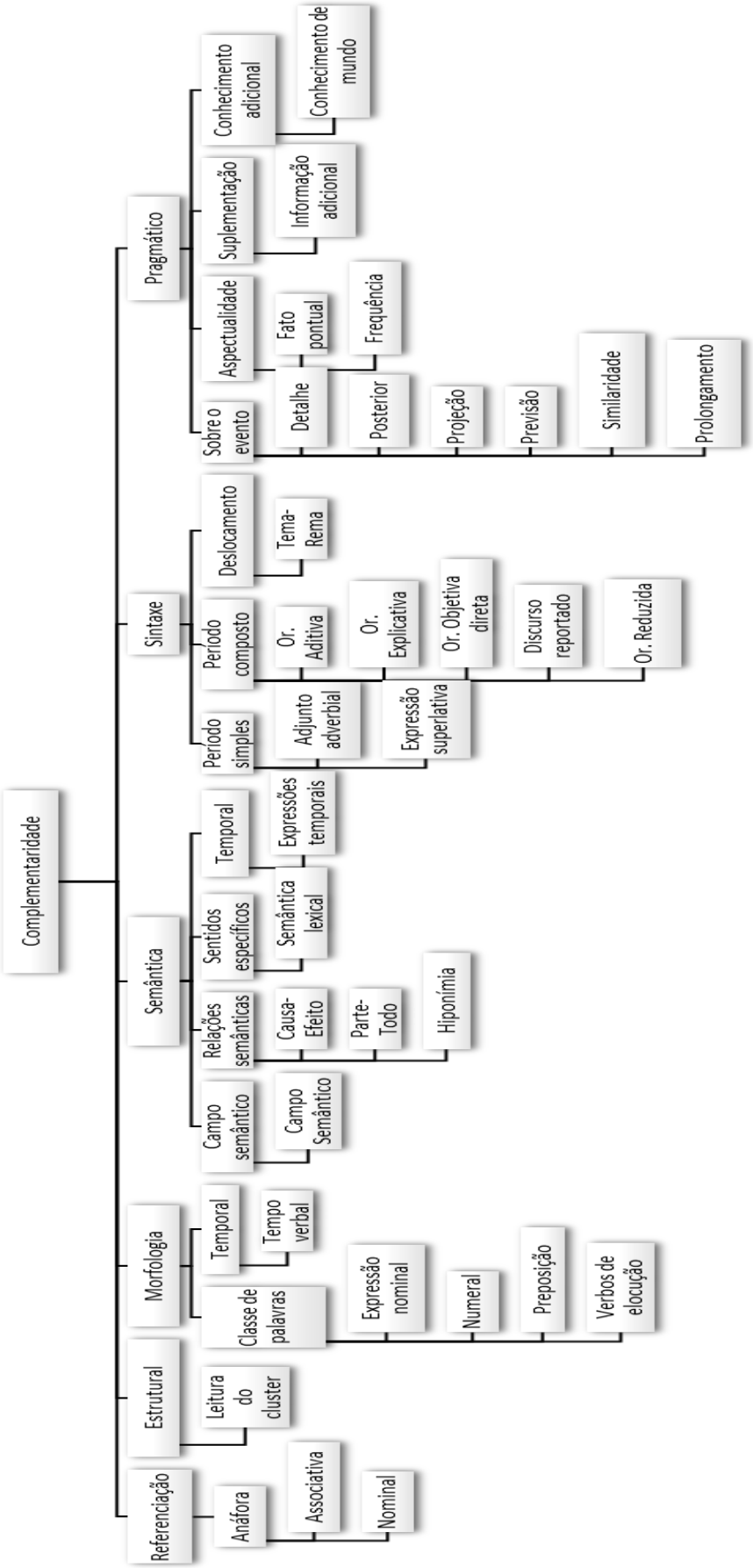

Figura 5: Tipologia dos dispositivos de sinalização da complementaridade.
Fonte: Elaborado pelo autor.

Na Figura 5, ilustra-se a tipologia dos dispositivos de sinalização que ocorrem em cada uma das categorias de análise linguística (p.ex. Morfológico e Sintático), suborganizados em sinais genéricos (p.ex. classe de palavras e período simples), os quais superordenam os sinais específicos (p.ex. numeral e adjunto adverbial).

Na Tabela 9, demonstra-se a quantificação de cada sinalizador em função das relações CST de complementaridade.

TIPOLOGIA			**RELAÇÃO CST DE COMPLEMENTARIDADE**			***TOTAL***
			ELABORATION	***FOLLOW-UP***	***HISTORICAL BACKGROUND***	
CATEGORIA	**SINAL GENÉRICO**	**SIINAL ESPECÍFICO**				
REFERENCIAÇÃO	**Anáfora**	ANÁFORA ASSOCIATIVA	50	31	0	81
		ANÁFORA NOMINAL	132	89	20	241
--------	**ESTRUTURAL**	LEITURA DO CLUSTER	48	60	14	122
MORFOLÓGICO	**CLASSE DE PALAVRAS**	NUMERAL	11	35	2	48
		EXPRESSÃO NOMINAL	2	15	0	17
		EXPRESSÃO PREPOSICIONAL	7	0	17	24
	TEMPORAL	TEMPO VERBAL	12	134	3	149
	VERBOS DE ELOCUÇÃO	VERBOS DE ELOCUÇÃO	26	51	0	77
SINTÁTICO	**PERÍODO SIMPLES**	ADJUNTO ADVERBIAL	31	40	2	73
		EXPRESSÃO SUPERLATIVA	0	0	26	26
	PERÍODO COMPOSTO	DISCURSO REPORTADO	67	52	0	119
		ORAÇÃO ADITIVA	26	2	0	28
		ORAÇÃO EXPLICATIVA	37	5	7	49
		ORAÇÃO OBJETIVA DIRETA	22	7	0	29
		ORAÇÃO REDUZIDA	12	3	0	15
	DESLOCAMENTO	TEMA-REMA	108	1	2	111
SEMÂNTICO	**CAMPO SEMÂNTICO**	CAMPO SEMÂNTICO	29	34	0	63
	RELAÇÕES SEMÂNTICAS	CAUSA-EFEITO	12	23	0	35
		HIPONÍMIA	16	4	0	20
		PARTE-TODO	42	15	0	57
	TEMPORAL	EXPRESSÃO TEMPORAL	4	109	57	170
	SENTIDO DE ACRÉSCIMO	SEMÂNTICA LEXICAL	27	42	8	77
PRAGMÁTICO	**Sobre o evento**	DETALHE	103	59	0	162
		POSTERIOR	0	92	0	92
		PREVISÃO	0	17	0	17
		PROLONGAMENTO	0	57	0	57
		PROJEÇÃO	0	18	0	18
		SIMILARIDADE	0	0	39	39
	Argumentação	FOCO ARGUMENTATIVO	17	0	0	17
	Suplementação	INFORMAÇÃO ADICIONAL	52	0	0	52
	Aspectualidade	FATO PONTUAL	0	0	38	38
		FREQUÊNCIA	0	0	38	38

Conhecimento adicional	CONHECIMENTO DE MUNDO	5	28	14	47

Tabela 9: Quantificação de dispositivos de sinalização em função das relações de complementaridade.

Fonte: Elaborado pelo autor.

Com base nesses resultados expressos na Tabela 9, destaca-se que alguns sinalizadores ocorrem com baixa frequência ou não ocorrem em algumas relações complementares. Os sinalizadores de natureza semântica e/ou pragmática, por exemplo, são característicos da relação CST *Historical Background*, sendo que alguns não ocorrem e outros ocorrem com frequência baixa. Os sinalizadores da categoria pragmático também parecem caracterizar a relação *Historical Background*, uma vez que, nessa relação, as informações temporais estão atreladas a ocorrência do sinalizador genérico *eventualidade*, manifestando-se especificamente por meio dos sinais *Frequente* ou *Fato Pontual*. Por fim, sobre a categoria sintático, nos pares anotados com a relação *Historical Background* observou-se a presença de e*xpressão superlativa* (como em "o maior acidente da história do país") apenas nessa relação CST. Tais apontamentos indicam que a ausência de um sinalizador em uma das relações CST pode indicar que ele seja característico de outra relação de complementaridade.

Ainda sobre a Tabela 9, percebe-se que há categorias de sinalizadores que ocorrem com maior frequência no *corpus* CSTNews, como pode-se observar na Figura 6.

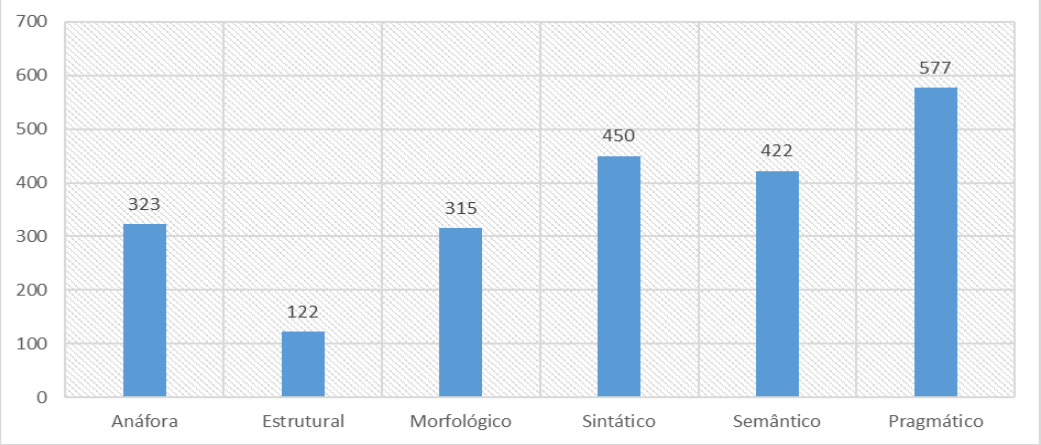

Figura 6: Quantificação dos dispositivos genéricos de sinalização da complementaridade.
Fonte: Elaborado pelo autor.

Na Figura 6, observa-se que os sinais dos níveis linguísticos sintático, semântico e pragmático têm destaque na caracterização manual realizada em todos os pares de sentenças do *corpus* CSTNews anotados com as três relações de complementaridade. Tal fato corrobora o fato de que as relações CST são de natureza semântica, inclusive as de caráter complementar. Assim,

considerar a classificação das relações que traduz o fenômeno na realização textual de informações temporais, apenas, não indica a totalidade da expressividade das relações.

No Quadro 10, exemplifica-se a ocorrência de cada um dos sinalizadores, os quais, quando possível, estão destacados, em função da relação CST em que o sinalizador ocorreu.

TIPOLOGIA			**EXEMPLOS**
CATEGORIA	**SINAL GENÉRICO**	**SINAL ESPECÍFICO**	
REFERENCIAÇÃO	**ANÁFORA**	ANÁFORA ASSOCIATIVA	RELAÇÃO: *ELABORATION* S1: Os corpos de **11 homens e 4 mulheres** de etnia tâmil, vestidos com camisetas da ACF, apareceram na semana passada no escritório da organização. S2: **Os funcionários da organização**, da etnia tâmil, trabalhavam com ajuda humanitária e reconstrução pós-tsunami.
		ANÁFORA NOMINAL	RELAÇÃO: *FOLLOW-UP* S1: Confrontos entre o Exército e o grupo rebelde Tigres Tâmeis eclodiram na região de Muttur há duas semanas, após a guerrilha ter cortado o suprimento de água para alguns vilarejos. S2: Os rebeldes afirmaram que consideram o novo bombardeio do Exército equivalente a "uma declaração de guerra".
-------	**ESTRUTURAL**	LEITURA DO CLUSTER	RELAÇÃO: *FOLLOW-UP* S1: Os ataques ocorreram apesar de um acordo assinado pelos Tigres Tâmeis para permitir a reabertura do reservatório de água e a ameaça de que novos bombardeios seriam considerados pelo grupo como uma declaração de guerra. S2: Os rebeldes afirmaram que consideram o novo bombardeio do Exército equivalente a "uma declaração de guerra".
MORFOLÓGICO	**CLASSE DE PALAVRAS**	NUMERAL	RELAÇÃO: *ELABORATION* S1: As seleções de vôlei e futebol conquistaram a Liga Mundial e a Copa América e escreveram mais uma vez o nome do Brasil nos respectivos esportes. S2: O Brasil é a melhor seleção das américas pela oitava vez.
		EXPRESSÃO NOMINAL	RELAÇÃO: *FOLLOW-UP* S1: Confrontos entre o Exército e o grupo rebelde Tigres Tâmeis eclodiram na região de Muttur há duas semanas, após a guerrilha ter cortado o suprimento de água para alguns vilarejos. S2: Os rebeldes afirmaram que consideram o **novo bombardeio** do Exército equivalente a "uma declaração de guerra".
		PREPOSIÇÃO	RELAÇÃO: *ELABORATION* S1: A quadrilha é acusada de praticar diversos crimes administrativos, como desvio de recursos públicos, corrupção, prevaricação, concussão, peculato, extorsão, lavagem de dinheiro e venda de sentenças judiciais.

		VERBOS DE ELOCUÇÃO	S2: Os recursos públicos eram desviados **para** pagamentos de serviços, compras e obras supostamente superfaturadas. RELAÇÃO: *FOLLOW-UP* S1: Fontes do Ministério de Transporte disseram que vários vagões descarrilaram e tombaram, e que os bombeiros conseguiram controlar um incêndio no trem que procedia de Lardo. S2: Fontes do Ministério de Transporte **disseram** que o número de mortos pode aumentar devido ao estado grave de vários dos feridos e que ainda há cadáveres sob os vagões dos dois trens.
	TEMPORAL	TEMPO VERBAL	RELAÇÃO: *FOLLOW-UP* S1: No momento, apenas um taxista e um passageiro **estavam** na rua. S2: A rua foi interditada e a perícia **está fazendo** a vistoria no local.
SINTÁTICO	**PERÍODO SIMPLES**	ADJUNTO ADVERBIAL	RELAÇÃO: *ELABORATION* S1: Como estratégia de obstrução, os partidos de oposição podem se utilizar, por mais de uma vez, de oito tipos de requerimentos para "travar" a votação, sem contar com comunicações de líderes e outros instrumentos que estendem a sessão, todos previstos no regimento. S2: Outro requerimento apresentado pelo DEM também pede o adiamento da votação por nove sessões.
		EXPRESSÃO SUPERLATIVA	RELAÇÃO: *HISTORICAL BACKGROUND* S1: Um ataque em dois lugares da Universidade Técnica da Virgínia, em Blacksburg, Estados Unidos, resultou na morte de 30 pessoas. S2: **O tiroteio é um dos piores crimes do tipo no *campus*** de uma universidade nos EUA desde que Charles Whitman abriu fogo do alto de uma torre no meio do campus da Universidade do Texas, em Austin, no dia 1º de agosto de 1966.
	PERÍODO COMPOSTO	DISCURSO REPORTADO	RELAÇÃO: *ELABORATION* S1: No domingo, o Exército bombardeou posições rebeldes na área. S2: Os rebeldes **afirmaram que consideram o novo bombardeio do Exército equivalente a "uma declaração de guerra**".
		ORAÇÃO ADITIVA	RELAÇÃO: *ELABORATION* S1: Nas duas primeiras etapas da obra, não será necessário o fechamento da pista. S2: Na segunda etapa, a parte concluída será reaberta **e a obra passará a ser feita na outra cabeceira**.

		ORAÇÃO EXPLICATIVA	RELAÇÃO: *ELABORATION* S1: Pelo menos 80 pessoas morreram e mais de 165 ficaram feridas nesta segunda-feira após a colisão de dois trens de passageiros no delta do Nilo, ao norte do Cairo, informaram fontes policiais e médicas. S2: Fontes do Ministério de Transporte disseram que o número de mortos pode aumentar **devido ao estado grave de vários dos feridos** e que ainda há cadáveres sob os vagões dos dois trens.
		ORAÇÃO OBJETIVA DIRETA	RELAÇÃO: *FOLLOW-UP* S1: Quinze funcionários locais de uma organização de caridade francesa no Sri Lanka foram encontrados mortos na cidade de Muttur, no norte do país. S2: O diretor da ACF no Sri Lanka, Benoit Miribel, confirmou a morte de seus funcionários e afirmou, comovido, **que a ONG "não sofreu uma perda similar em seus mais de 25 anos de existência".**
		ORAÇÃO REDUZIDA	RELAÇÃO: *ELABORATION* S1: Nas duas primeiras etapas da obra, não será necessário o fechamento da pista. S2: Na primeira etapa, será reformado um terço da pista, em uma das cabeceiras, **ficando o restante disponível para pousos e decolagens**.
	DESLOCAMENTO	TEMA-REMA	RELAÇÃO: *ELABORATION* S1: De acordo com a polícia da capital russa, uma bomba causou o **incidente**. S2: **A explosão**, cujas causas ainda são desconhecidas, aconteceu às 10h40 (3h40 em Brasília) no mercado Cherkizov, localizado no nordeste da capital russa.
SEMÂNTICO	**CAMPO SEMÂNTICO**	CAMPO SEMÂNTICO	RELAÇÃO: *ELABORATION* S1: Os rebeldes afirmaram que consideram o novo bombardeio do Exército equivalente a "uma declaração de guerra". S2: Os **ataques** ocorreram apesar de um acordo assinado pelos Tigres Tâmeis para permitir a reabertura do reservatório de água e a **ameaça** de que novos **bombardeios** seriam considerados pelo grupo como uma declaração de **guerra**.
	RELAÇÕES SEMÂNTICAS	CAUSA-EFEITO	RELAÇÃO: *ELABORATION* S1: O CGE (Centro de Gerenciamento de Emergências) da Prefeitura de São Paulo registrava oito pontos de **alagamento** na cidade, às 9h30 desta segunda-feira. S2: Segundo o Centro de Gerenciamento de Emergências (CGE), **a última chuva havia ocorrido no dia 29 de junho, em intensidade menor do que a que cai sobre a cidade desde a noite de domingo**.

		HIPONÍMIA	RELAÇÃO: *ELABORATION* S1: Segundo divulgada pela PF, **o grupo criminoso** desviou desde 2004 cerca de R$ 70 milhões dos cofres públicos, por meio do pagamento de serviços, compras e obras superfaturadas. S2: **Dos 24 deputados estaduais, 23** praticavam esse crime e, segundo a PF, desviaram mais de R$ 10 milhões com essa prática, em apenas um ano.
		PARTE-TODO	RELAÇÃO: *FOLLOW-UP* S1: **A pista auxiliar de Congonhas** abriu às 6h, apenas para decolagens. S2: **Congonhas** só abriu para pousos, às 8h50.
	TEMPORAL	EXPRESSÃO TEMPORAL	RELAÇÃO: *HISTORICAL BACKGROUND* S1: No caso do Japão, a magnitude apontada de 6,8 é considerada "forte". S2: Foi o pior do país **desde 1995**, quando um tremor de magnitude 7,3 matou mais de 6.400 pessoas na cidade de Kobe.
	SENTIDO DE ACRÉSCIMO	SEMÂNTICA LEXICAL	RELAÇÃO: *FOLLOW-UP* S1: "Tentamos enviar uma equipe a Muttur para averiguar o que está acontecendo, mas os soldados não permitiram que entrássemos na cidade, que está totalmente bloqueada", afirmou. S2: Até o momento, as autoridades do Sri Lanka **não confirmaram** as mortes ou **esclareceram** o que acontece na cidade de Muttur.
PRAGMÁTICO	**SOBRE O EVENTO**	DETALHE	RELAÇÃO: *ELABORATION* S1: Pelo menos oito ataques foram confirmados. S2: Mais de dez agências bancárias, um posto de gasolina e um supermercado foram atacados.
		POSTERIOR	RELAÇÃO: *FOLLOW-UP* S1: Os votos brancos e nulos somam 9%, e 9% dos entrevistados não opinaram. S2: **Os votos brancos e nulos somaram 10% e aqueles que não sabem ou não opinaram são 4%**.
		PREVISÃO	Relação: *Follow-up* S1: Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril. S2: **O revezamento terminará em 8 de agosto**, primeiro dia das Olimpíadas de Pequim.
		PROLONGAMENTO	RELAÇÃO: *FOLLOW-UP* S1: Um dia antes do acidente, na segunda-feira, 16, o avião também teria apresentado problemas ao aterrissar em Congonhas, durante o voo 3215, procedente de Belo Horizonte (Confins), só conseguindo parar muito próximo do final da pista. S2: O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, **e ainda assim a aeronave da TAM,**

		um Airbus A320, continuou voando, com o reverso direito desligado.
	PROJEÇÃO	RELAÇÃO: *FOLLOW-UP* S1: Em junho, o presidente tinha 53% das intenções de voto contra 29% do tucano. S2: **De acordo com a pesquisa, Lula (PT) tem 44% das intenções de voto, contra 25% de Geraldo Alckmin (PSDB) e 11% de Heloísa Helena (PSOL)**.
	SIMILARIDADE	RELAÇÃO: *HISTORICAL BACKGROUND* S1: No domingo, o LTTE ofereceu o cessar-fogo em troca do desbloqueio da presa, mas o Governo rejeitou a oferta e lançou uma nova ofensiva sobre a área, de onde mais de 15 mil civis foram obrigados a fugir na última semana. S2: **Os confrontos em Muttur têm sido os mais intensos na ilha desde a assinatura de um cessar-fogo há quatro anos**.
ARGUMENTAÇÃO	FOCO ARGUMENTATIVO	RELAÇÃO: *ELABORATION* S1: "O reverso é um instrumento auxiliar na hora de frear o avião, mas não é considerado fator importante para o pouso, portanto, mesmo não funcionando, ele poderia ter brecado a aeronave sem problemas", diz. S2: O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, e ainda assim a aeronave da TAM, um Airbus A320, continuou voando, com o reverso direito desligado.
SUPLEMENTAÇÃO	INFORMAÇÃO ADICIONAL	RELAÇÃO: *ELABORATION* S1: Mas, diante da dificuldade para encontrar pessoas que aceitassem assumir uma das diretorias da agência reguladora, após a renúncia de três diretores, Jobim decidiu indicar a economista para o cargo. S2: **Como os diretores de agências têm mandato de cinco anos**, só podem sair por renúncia, decisão judicial ou acusação de improbidade administrativa.
ASPECTUALIDADE	FATO PONTUAL	RELAÇÃO: *HISTORICAL BACKGROUND* S1: O congestionamento esteve ainda maior às 9h, quando chegou a 113 km de extensão para uma média de 32 km. S2: **Em julho do ano passado**, a média foi de 36 km no horário.
	FREQUÊNCIA	RELAÇÃO: *HISTORICAL BACKGROUND* 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.

CONHECIMENTO ADICIONAL	CONHECIMENTO DE MUNDO	RELAÇÃO: *ELABORATION* S1: A oposição passará o dia tentando obstruir os trabalhos em plenário com o único objetivo de retardar a votação e dificultar a tarefa governista. S2: **A Câmara discute o requerimento de autoria do PSDB para adiamento da discussão da proposta que prorroga a CPMF**.

Quadro 10: Exemplos de ocorrência dos dispositivos de sinalização da complementaridade.
Fonte: Elaborado pelo autor.

Como apontado nas análises anteriores, nem sempre uma relação CST de complementaridade caracteriza-se pela presença de apenas um sinalizador. Os exemplos que constam no Quadro 10, dispostos de maneiras categóricas evidenciando categoria, sinais genéricos e específicos e a relação CST, deve ter essa consideração como ressalva. Em "*Fontes do Ministério de Transporte disseram que o número de mortos pode aumentar devido ao estado grave de vários dos feridos e que ainda há cadáveres sob os vagões dos dois trens*" a complementaridade é manifesta por meio dos sinais verbo de elocução (no caso, *disseram*), discurso reportado e oração objetiva direta (que, no caso, coincidem com *que o número de mortos pode aumentar*).

Além disso, há sinalizadores semelhantes entre si, em especial da categoria Pragmático (por exemplo *Projeção* e *Previsão*). Assim, apresenta-se o Quadro 11, que traz as definições de cada um dos sinalizadores e de suas especificidades.

TIPOLOGIA			**DEFINIÇÃO**
CATEGORIA	**SINAL GENÉRICO**	**SINAL ESPECÍFICO**	
REFERENCIAÇÃO	**ANÁFORA**	ANÁFORA ASSOCIATIVA	Em um par de sentenças (S1,S2) os referentes principais da narrativa de S1 são retomados e elaborados em S2 sob a forma de associação (*como "o maior corpo celeste" e "o novo membro"*).
		ANÁFORA NOMINAL	Em um par de sentenças (S1,S2) os referentes da narrativa de S1 são retomados em S2, podendo descrever eventos que se sucedem ou são previstos a partir dele.
-------	**ESTRUTURAL**	LEITURA DO CLUSTER	Em um par de sentenças (S1,S2) em que a delimitação da complementaridade esteja comprometida, tal sinal auxilia a compreensão do fenômeno.
MORFOLÓGICO	**CLASSE DE PALAVRAS**	NUMERAL	Em um par de sentenças (S1,S2) os numerais podem indicar em S2 a atualização do evento narrado em S1, em que o evento ou os referentes ligados a ele (como "*candidatos à presidência*") são retomados por meio de alguma manobra anafórica.
		EXPRESSÃO NOMINAL	Em um par de sentenças (S1,S2), S2 apresenta uma expressão nominal que caracteriza-se por ser uma informação adicional que pode, inclusive, indicar tempo (*como "funcionária de*

			carreira").
		PREPOSIÇÃO	Em um par de sentenças (S1,S2), S2 apresenta preposição ou locução prepositiva que indica tempo (como *durante*) ou finalidade (como *para*).
		VERBOS DE ELOCUÇÃO	Em um par de sentenças (S1,S2), esse sinal evidencia a presença de discurso reportado em S2 (*como "disse" e "afirmou"*), marcando a elaboração de algum ponto da narrativa de S1.
	TEMPORAL	TEMPO VERBAL	Em um par de sentenças (S1,S2), esse sinalizador ocorre em S2 ao demonstrar em sua desinência verbal tempo-modo o tempo do evento narrado em relação a S1; salienta-se que esse dispositivo mantém contraste entre os verbos de S1 e S2.
SINTÁTICO	**PERÍODO SIMPLES**	ADJUNTO ADVERBIAL	Em um par de sentenças (S1,S2), esse sinal evidencia em S2 detalhes sobre o evento já narrado em S1, como informações de lugar (como *em São Domingos*) ou algum advérbio que evidencie acréscimo (como *também*).
		EXPRESSÃO SUPERLATIVA	Em um par de sentenças (S1,S2), esse sinal evidencia um evento em S2 que, em relação a S1, nunca obtivera as proporções descritas.
	PERÍODO COMPOSTO	DISCURSO REPORTADO	Em um par de sentenças (S1,S2), esse dispositivo de sinalização cumpre a função de trazer detalhes em S2 sobre o evento narrado em S1, sempre evidenciando a fonte do discurso.
		ORAÇÃO ADITIVA	Em um par de sentenças (S1,S2), esse sinal evidencia, em S2, o acréscimo de informações por meio de uma oração aditiva, sempre introduzida por uma conjunção aditiva.
		ORAÇÃO EXPLICATIVA	Em um par de sentenças (S1,S2), esse sinal evidencia, em S2, explicações de fatos narrados em S1 por meio de uma oração explicativa.
		ORAÇÃO OBJETIVA DIRETA	Em um par de sentenças (S1,S2), esse dispositivo de sinalização aponta o trecho que compreende a informação complementar em S2, uma vez que o referente de S1, por manobra anafórica, passa a ocupar lugar de tópico/sujeito da sentença.
		ORAÇÃO REDUZIDA	Em um par de sentenças (S1,S2), esse sinal evidencia, em S2, o acréscimo de informações por meio de uma oração reduzida de gerúndio, particípio ou infinitivo.
	DESLOCAMENTO	TEMA-REMA	Em um par de sentenças (S1,S2), o Rema de S1 é retomado em S2 sob a forma de Tema, caracterizando sua elaboração por meio desse deslocamento.
SEMÂNTICO	**CAMPO SEMÂNTICO**	CAMPO SEMÂNTICO	Em um par de sentenças (S1,S2), esse sinalizador opera de forma a auxiliar a compreender a sucessão do evento de S1 em S2, em que o evento seja retomado por alguma manobra anafórica.
	RELAÇÕES SEMÂNTICAS	CAUSA-EFEITO	Em um par de sentenças (S1,S2), S2 apresenta ora a causa, ora o efeito do evento narrado em S1.
		HIPONÍMIA	Em um par de sentenças (S1,S2), S2 apresenta a especificação de algum referente ou informação que esteja genérica em S1.
		PARTE-TODO	Em um par de sentenças (S1,S2), S2 apresenta referentes que integram outros que estejam sendo narrados em S1.
	TEMPORAL	EXPRESSÃO TEMPORAL	Em um par de sentenças (S1,S2), esse sinalizador ocorre em S2 ao narrar um evento passado/histórico/projetado/previsto em relação a S1 sob a forma de contraste entre informações temporais.

	SENTIDO DE ACRÉSCIMO	SEMÂNTICA LEXICAL	Em um par de sentenças (S1,S2), S2 apresenta alguma unidade lexical cujo sentido é de acrescentar detalhe a algum ponto da narrativa de S1, (*como "ressaltar", "destacar" e "completar"*).
PRAGMÁTICO	**SOBRE O EVENTO**	DETALHE	Em um par de sentenças (S1,S2) em que narra-se sobre o mesmo evento, S2 apresenta algum detalhe sobre o evento narrado em S1, em que tal informação esteja aconrada no fato de S2 ter já acontecido após S1 (*como é o caso de atualização de informação*).
		POSTERIOR	Em um par de sentenças (S1,S2), S2 narra algum evento diferente que aconteça depois de S1.
		PREVISÃO	Em um par de sentenças (S1,S2), S2 narra algum evento que possivelmente aconteça depois de S1.
		PROLONGAMENTO	Em um par de sentenças (S1,S2), S2 narra o prolongamento do evento narrado em S1.
		PROJEÇÃO	Em um par de sentenças (S1,S2), S2 narra algum evento que certamente aconteça depois de S1.
		SIMILARIDADE	Em um par de sentenças (S1,S2), S2 apresenta um evento diferente de S1, mas que sejam semelhantes em sua natureza eventiva (como acidentes aéreos).
	ARGUMENTAÇÃO	FOCO ARGUMENTATIVO	Em um par de sentenças (S1,S2), S2 apresenta outra perspectiva daquilo que é o foco de S1, por vezes sobe a forma de justificativa.
	SUPLEMENTAÇÃO	INFORMAÇÃO ADICIONAL	Em um par de sentenças (S1,S2), S2 apresenta uma informação adicional não prevista em S1, já que não faz parte de seu foco informacional.
	ASPECTUALIDADE	FATO PONTUAL	Em um par de sentenças (S1,S2), S2 narra sobre um evento que tenha acontecido apenas uma vez no passado, em relação a S1.
		FREQUÊNCIA	Em um par de sentenças (S1,S2), S2 narra sobre um evento que tenha acontecido mais de uma vez no passado ou que venha acontecendo, em relação a S1.
	CONHECIMENTO ADICIONAL	CONHECIMENTO DE MUNDO	Em um par de sentenças (S1,S2), esse sinalizador opera de forma a auxiliar a compreender a sucessão do evento de S1 em S2, uma vez que só será possível após apreender que os referentes utilizados nas sentenças do par possuem relação entre si apenas fora do texto (como *CET e São Paulo*).

Quadro 11: Definição dos dispositivos de sinalização da complementaridade.
Fonte: Elaborado pelo autor

Como visto na Tabela 9, há dispositivos de sinalização que não ocorrem nos pares de sentenças anotados com certas relações de complementaridade, uma vez que possuem ocorrência zero. Tal informação é relevante já que os algoritmos de AM têm melhor desempenho quando lidam com sinalizadores específicos de cada relação. Além disso, tal análise dá subsídio para interpretar informações advindas dos classificadores automáticos (discussão que será retomada no próximo capítulo)

No Quadro 12[25], demonstram-se os dispositivos de sinalização específicos de cada uma das relações.

Dispositivo			**Relação CST**		
			Elaboration	*Foloow-up*	*Hisotrical background*
Categoria	Sinal genérico	Sinal específico			
Referenciação	Anáfora	Anáfora associativa	X	X	NSA
		Anáfora nominal	X	X	X
------	Estrutural	Leitura do cluster	X	X	X
Morfologia	Classe de palavras	Numeral	X	X	NSA
		Expressão nominal	NSA	NSA	X
		Preposição	X	NSA	X
		Verbos de elocução	X	X	NSA
	Temporal	Tempo verbal	X	X	NSA
Sintaxe	Período simples	Adjunto adverbial	X	X	X
		Expressão superlativa	NSA	NSA	X
	Período composto	Oração aditiva	X	X	X
		Oração explicativa	X	X	X
		Oração objetiva direta	X	X	NSA
		Discurso reportado	X	X	NSA
		Oração reduzida	X	NSA	NSA
	Deslocamento	Tema-Rema	X	NSA	NSA
Semântica	Campo semântico	Campo Semântico	X	X	NSA
	Relações semânticas	Causa-Efeito	X	X	NSA
		Parte-Todo	X	X	NSA
		Hiponímia	X	NSA	NSA
	Temporal	Expressão temporal	X	X	X
	Sentido de acréscimo	Semântica lexical	X	X	X
Pragmática	Sobre o evento	Detalhe	X	X	NSA
		Posterior	NSA	X	NSA
		Projeção	NSA	X	NSA
		Previsão	NSA	X	NSA
		Similaridade	NSA	NSA	X
		Prolongamento	NSA	X	NSA
	Aspectualidade	Fato pontual	NSA	NSA	X
		Frequência	NSA	NSA	X

[25] No Quadro 12, quando houver "NSA" deve interpretar que aquele atributo não é característico daquela relação e Não Se Aplica à análise; quando houver "X" é porque aquele atributo ocorreu em pares de sentenças anotados com dada relação CST de complementaridade.

SUPLEMENTAÇÃO	INFORMAÇÃO ADICIONAL	X	NSA	NSA
ARGUMENTAÇÃO	FOCO ARGUMENTATIVO	X	NSA	NSA
CONHECIMENTO ADICIONAL	CONHECIMENTO DE MUNDO	X	X	X

Quadro 12: Ocorrência dos sinalizadores em função das relações CST de complementaridade.
Fonte: Elaborado pelo autor.

No Quadro 12, tem-se a ocorrência de cada um dos sinalizadores em função das relações CST que traduzem a complementaridade. O dispositivo *Anáfora nominal*, por exemplo, ocorre nos pares com as relações *Elaboration*, *Follow-up* e *Historical background*, enquanto o dispositivo *Anáfora associativa* não ocorre nos pares anotados com a relação *Historical background*, sendo indicado por NSA.

Como apresentado anteriormente, ao realizar a anotação do CSTNews, Aleixo e Pardo (2008b) adaptaram o conjunto de relações CST, agrupando algumas relações sob um único rótulo. As relações de *Description* e *Reader profile*, por exemplo, foram interpretadas como *Elaboration*, e *Fulfillment* foi agregada à *Follow-up*. Assim, uma das dificuldades da caracterização dos pares de sentenças foi deparar-se diante de possíveis subclassificações para as relações *Follow-up* e *Elaboration* que eram resultantes das adaptações realizadas pelo agrupamento de relações semelhantes em um único rótulo.

A decisão de unificar alguns rótulos de relações ainda resultou na limitação da organização tipológica das relações CST de complementaridade, a saber, temporal e atemporal (MAZIERO *et al.*, 2010). Com base no estudo e na descrição manual realizados, aponta-se uma série de sinalizadores que caracterizam a complementaridade para além daqueles de natureza puramente temporal (sinalizadores de natureza semântica e pragmática, por exemplo). Dessa maneira, a organização tipológica proposta, até então, é limitada.

Diante da inviabilidade de reanotar os pares de sentenças, os sinalizadores do tipo pragmático foram propostos com o objetivo de capturar diferentes nuances ou interpretações específicas de cada uma das relações de complementaridade.

Baseando-se nas análises do *corpus*, propõe-se o refinamento das definições das relações CST de complementaridade propostas inicialmente por Maziero *et al.* (2010).

a) *Elaboration*: dado o par de sentenças (S1 e S2), S2 detalha/refina/elabora/descreve algum elemento presente em S1 por meio de focalização e/ou topicalização frasal, sendo que S2 não deve repetir informações presentes em S1;

b) *Follow-up*: dado o par de sentenças (S1 e S2), S2 pode apresentar eventos subsequentes aos narrados em S1 ou mesmo prever eventos em decorrência dos veiculados por S1;

c) *Historical background:* dado o par de sentenças, S1 e S2, S2 apresenta informações históricas de caráter frequente ou pontual sobre algum elemento presente em S1, sendo que os eventos de S1 e S2 devem ser da mesma natureza.

No próximo capítulo, apresentam-se os testes e a avaliação dos classificadores criados a partir dos dispositivos de sinalização da tipologia apresentada neste capítulo.

Capítulo 5

TESTE E AVALIAÇÃO DOS CLASSIFICADORES DE IDENTIFICAÇÃO DA COMPLEMENTARIDADE

Neste capítulo, procuro apresentar conceitos basilares do Aprendizado de Máquina. Além disso, faço uma apresentação mais detalhada sobre a ferramenta Weka, utilizada para a geração dos classificadores das relações CST de complementaridade neste trabalho. Por fim, apresento o treinamento dos classificadores e avaliação.

5.1. Aprendizado de Máquina

De acordo com Monard e Baranaukas (2003), o Aprendizado de Máquina (AM), como subárea da Inteligência Artificial, visa ao desenvolvimento e aprimoramento de técnicas computacionais, bem como a criação de sistemas de aquisição de conhecimento de maneira automatizada. Tais sistemas são capazes de tomar decisões baseadas em experiências acumuladas previamente por meio da resolução bem-sucedida de problemas anteriores.

De acordo com Mitchell (1997), os sistemas de AM podem ser utilizados em paradigmas *estatísticos* (paradigmas baseados em modelos estatísticos que se aproximam do conceito induzido), *conexionista* (que se caracterizam pelas redes neurais matemáticas), *instance-based* (paradigma cujos modelos são baseados em exemplos), *genético* (cuja predição é baseada em uma população de informações) e *simbólico*. O paradigma simbólico, em especial, é mais utilizado em tarefas que interpretação humana, sendo representado por árvores de decisão ou por conjuntos de regras. Entretanto, é necessário admitir que não há como saber previamente qual paradigma e/ou algoritmo terá melhor desempenho, uma vez que sua avaliação está atrelada à tarefa a ser desenvolvida (PRATI *et al.*, 2001).

Assim, Carbonell *et al.* (1983) classifica os sistemas de AM em (i) caixa-preta, cujas interpretações internas podem não ser facilmente legíveis pelo humano, e (ii) orientados a conhecimento, cujas estruturas simbólicas são interpretáveis ao humano.

Em tarefas de descrição linguística para o PLN, o AM tem demonstrado diversas vantagens teórico-metodológicas. Antunes *et al.* (2017), ao descrever a formação de gentílicos em PB, argumenta que é possível, por meio do AM, verificar regularidades linguísticas (no caso, morfológicas), além de identificar e extrair eventuais padrões interessantes à tarefa em questão. Os autores ainda salientam que, caso o AM apresente boa acurácia, pode confirmar hipóteses lançadas de maneira manual, além de identificar certos padrões que a reflexão humana pode não identificar, ainda que não alcance os mesmos níveis de cognição.

Assim, com o objetivo de extrair automaticamente conhecimentos antes desconhecidos da análise puramente humana/manual das relações CST de complementaridade, apresenta-se o *software* Weka na próxima seção, o qual disponibiliza diversos algoritmos automáticos capazes de construir classificadores automáticos.

5.2. Waikato Environment for Knowledge Analysis

O *Waikato Environment for Knowledge Analysis* (Weka) (WITTEN, FRANK, 2005) é um ambiente que engloba um conjunto de algoritmos de AM. O Weka contém ferramentas para preparação de dados, classificação, regressão, agrupamento, mineração de regras de associação e visualização. Nesta seção, destacam-se as ferramentas que auxiliam na classificação, avaliação e visualização dos classificadores automáticos[26].

5.2.1. Preparação dos arquivos

De modo geral, os arquivos que serão submetidos ao Weka não podem apresentar nenhum relacionamento referencial ou explícito entre os dados e as instâncias, isto é, as instâncias não podem já estar pré-definidas em relação aos dados/características. Os arquivos deverão estar em formatos legíveis pelo *software*. Como *default*, o sistema utiliza o padrão *Attribute-*

[26] De acordo com Prati *et al.* (2001), um classificador automático seria uma generalização de exemplos fornecidos ao sistema de AM, pois, a partir de um conjunto pré-definido, o sistema é capaz de classificar automaticamente outros exemplos com base em suas experiências anteriores.

Relation File Format (ARFF), mas o Weka também aceita o formato *Comma Separated Values* (CSV).

No formato ARFF, os arquivos devem ter 3 características básicas: (i) cabeçalho (no exemplo abaixo, "@RELATION"), (ii) declaração de atributo ("@ATTRIBUTE") e (iii) dados da seção ("@DATA"). Em (i), declara-se o nome do arquivo; em (ii), declaram-se os atributos que compõem a instância; em (iii), os dados que compõem as instâncias, separados por vírgulas, como exemplificado no Quadro 13.

```
@RELATION ARQUIVO_EXEMPLO_TESTE

@ATTRIBUTE PAR NUMERIC
@ATTRIBUTE RELACAO {ELABORATION, FOLLOW_UP, HISTORICAL_BACKGROUND}
@ATTRIBUTE ANAFORA {ASSOCIATIVA, NOMINAL, NENHUM}
@ATTRIBUTE MORFOLOGICO {ADVERBIO_TEMPORAL, NUMERAL, PREPOSICAO, NENHUM}
@ATTRIBUTE TEMPORAL {EXPRESSAO_TEMPORAL, TEMPO_VERBAL, NENHUM}

@DATA
01,ELABORATION,ASSOCIATIVA,NENHUM, EXPRESSAO_TEMPORAL
02,ELABORATION,NENHUM,PREPOSICAO,NENHUM
03,ELABORATION,ASSOCIATIVA,PREPOSICAO,NENHUM
```

Quadro 13: Exemplo de construção de arquivo no formato ARRF para o Weka.
Fonte: Elaborado pelo autor.

No Quadro 13, tem-se um exemplo de descrição da relação CST *Elaboration*. O conjunto de análise é composto por três instâncias, ou seja, pares de sentenças. O par 01, por exemplo, foi anotado com a relação *Elaboration*, a qual foi caracterizada por apresentar *Anáfora associativa*, *nenhum* atributo *Morfológico*, além de apresentar e*xpressão temporal* como atributo *temporal*. Ao ser submetido a algum algoritmo de AM, poderia ser apontado que, dado o conjunto de 3 instâncias, a relação *Elaboration* poderia ser caracterizada por anáfora associativa, preposição e nenhum atributo temporal. Esse possível classificador teria uma cobertura de 0,66% dos casos, isto é, duas das três instâncias.

Já no formato CSV, os arquivos devem ter apenas os atributos e as instâncias. Muito comumente, a produção desses arquivos é feita via tabelas do Excel® (ou similares) e convertidas para o formato CSV, como no exemplo do Quadro 14.

PAR	RELACAO	ANAFORA	MORFOLOGICO	TEMPORAL
1	ELABORATION	ASSOCIATIVA	PREPOSICAO	EXPRESSAO_TEMPORAL
2	ELABORATION	NENHUM	NENHUM	NENHUM
3	ELABORATION	NENHUM	NENHUM	NENHUM

Quadro 14: Exemplo de análise em tabela para o formato CSV.
Fonte: Elaborado pelo autor.

```
PAR,RELACAO,ANAFORA,MORFOLOGICO,TEMPORAL
01,ELABORATION,ASSOCIATIVA,PREPOSICAO, EXPRESSAO_TEMPORAL
02,ELABORATION,NENHUM,NENHUM,NENHUM
03,ELABORATION,NENHUM,NENHUM,NENHUM
```

Quadro 15: Exemplo de construção de arquivo no formato CSV para o Weka.
Elaborado pelo autor.

No Quadro 15, tem-se a conversão da tabela Excel do Quadro 14 para o formato CSV. Especificamente, tem-se um exemplo de descrição da relação CST *Elaboration*. Como conjunto de análise, têm-se três instâncias (novamente pares de sentenças), com as informações de par, relação, anáfora, morfológico e temporal, respectivamente. No Par 02, por exemplo, anotado com a relação *Elaboration*, apresenta *nenhum* como preenchimento dos atributos Anáfora, Morfológico e Temporal. Assim, em um possível classificador, ter-se-ia que essa relação CST não é caracterizada por nenhum dos atributos descritos, tendo cobertura de 0.66%, isto é, dois dos três casos.

É importante ressaltar que a ordem com que são preenchidas as instâncias não deve ser diferente, para que o arquivo seja legível pelo sistema. O par 01, por exemplo, que apresenta os traços anáfora e morfológico, respectivamente, não deve diferir da ordem no preenchimento dos demais pares. Caso haja alguma complicação (por digitação, por exemplo) o *software* fará indicação, por linha, onde há o erro.

5.2.2. Geração de classificadores

De posse de um arquivo em um dos formatos legíveis pelo Weka, o usuário deverá submetê-lo a um dos classificadores disponíveis na aba *Classify* do *software*. Para construir uma árvore de decisão, por exemplo, o usuário poderá selecionar o algoritmo J48. A Figura 8 exemplifica uma possível árvore baseando-se nos dados do Quadro 14.

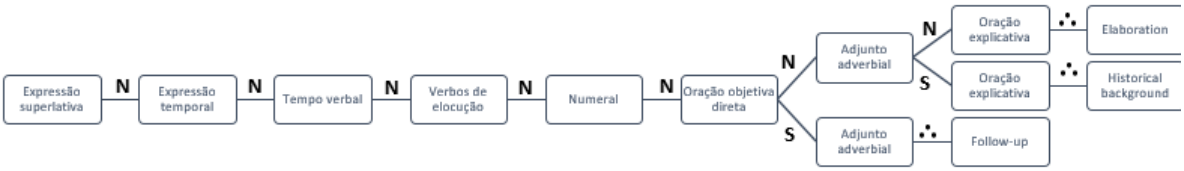

Figura 7: Exemplo de Árvore de Decisão.

Fonte: Elaborado pelo autor.

Na Figura 6 exemplifica-se o recorte de uma árvore de decisão gerada a partir do algoritmo J48. Nela, há atributos (expressos pelos valores categóricos/nominais em caixas) seguidos de seus valores ("N" para "não" e "S" para "sim") resultando em uma das três relações CST de complementaridade. Assim, tem-se que quando, por-exemplo, expressão superlativa, epressão temporal, tempo verbal, verbos de elocução. Numeral, oração objetiva direta, adjunto adverbial, oração explicativa obtiverem valor "N" a relação será *Elaboration*.

Na Figura 8, demonstra-se a tela principal do *software* Weka.

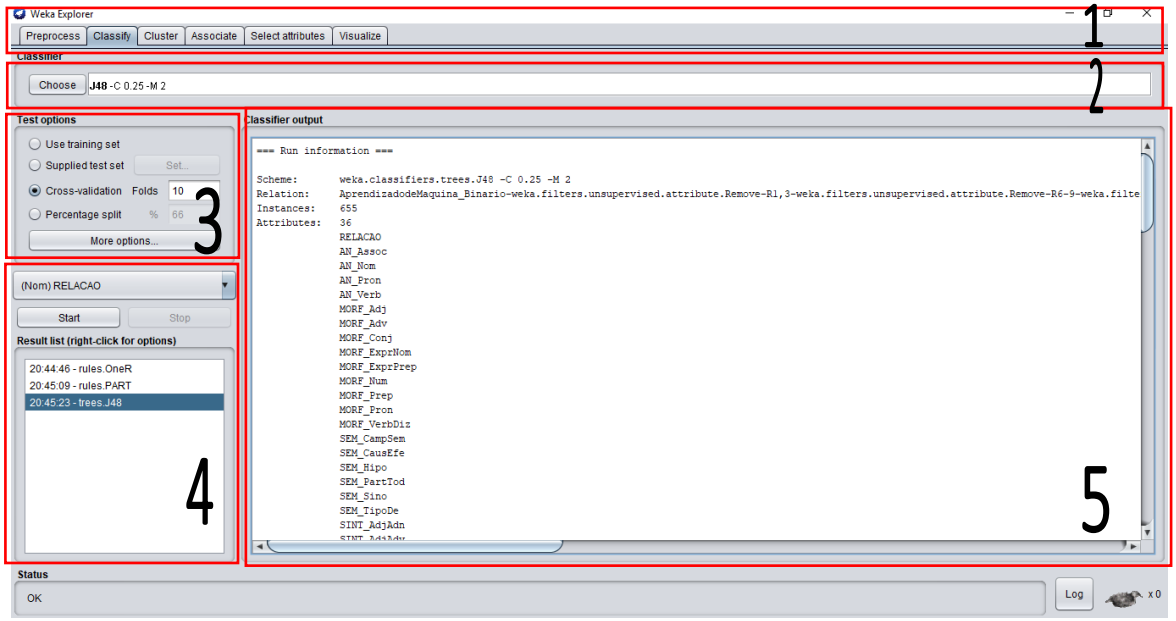

Figura 8: Tela principal do Weka.
Fonte: Fonte: Baseado em Witten e Frank (2005).

A Figura 8 demonstra a tela principal do Weka, que pode ser dividida em cinco seções distintas, demarcadas na própria figura.

Na Seção 1, encontram-se todas as abas de funcionalidades do software, a saber:

(i) *Process*: área em que o usuário faz o *upload* de seu conjunto de dados, bem como manipula-o (como excluir atributos que não serão utilizados na análise, p.ex.);

(ii) *Classify*: campo onde são criados e também avaliados os classificadores por meio de algoritmos automáticos;

(iii) *Cluster*: área onde se desenvolve e avalia automaticamente um modelo de conjunto de treinamento com base no conjunto de teste, submetido ao Weka;

(iv) *Associate*: área em que se identificam as melhores associações de atributos em pares;

(v) *Selected Atributes*: campo em que se verificam os atributos que têm melhor desempenho para a geração de um classificador. Tal etapa é detalhada na Subseção 5.2.3;

(vi) *Visualize*: área em que se identifica, em um eixo cartesiano, a funcionalidade dos atributos.

Na Seção 2, o usuário pode eleger o algoritmo automático que utilizará para produzir os classificadores, a depender do paradigma pretendido.

Na Seção 3, *Test Options*, o Weka oferece quatro opções para realizar os testes, a saber:

1. *Use training set*: Nessa opção, o usuário testará o conhecimento no mesmo conjunto de treinamento. Essa opção não é muito aconselhável, já que, no conjunto de teste, a hipótese (o conjunto do qual se apreendem as regras) demonstra baixa capacidade de generalização (ou classificação), resultando em *overfitting*. Outro aspecto negativo dessa opção é que a hipótese memorizou e/ou se especializou nos dados de treinamento.

2. *Supplied test set*: Nessa opção, o usuário concede ao sistema um conjunto de teste, o qual será distinto do conjunto de treinamento (já concedido ao *software* na tela inicial do programa). Assim, o algoritmo aprenderá hipóteses no conjunto de treinamento e generalizará o conhecimento no conjunto de teste. É importante frisar que os dois arquivos precisam ter a mesma quantidade de atributos.

3. *Cross-validation folds*: Nessa opção, o usuário cederá ao *software* somente um conjunto de análise de dados. Assim, os atributos são validados de maneira cruzada em uma quantidade predeterminada de pastas. Em uma análise cruzada de 10 pastas, por exemplo, particiona-se o conjunto de dados em outros 10 subconjuntos, em que, a cada execução do algoritmo escolhido, são utilizados nove subconjuntos para classificar (simulando o conjunto de treinamento) e somente um para validar o classificador (simulando o conjunto de teste). Assim, ao final, tem-se o resultado simulado em 10 pastas diferentes.

4. *Percentage Split*: Nessa opção, o usuário também cederá ao sistema somente um conjunto de dados, mas escolherá uma parte dele para ser o conjunto de treinamento e outra para ser de teste. A diferença da opção anterior é que o mesmo conjunto será testado somente uma vez, ao invés de 10, como no exemplo.

Na Seção 4, inicia-se a construção do classificador (clicando em *Start*) com base no atributo escolhido. Na Figura 8, para criar um classificador, escolheu-se, por exemplo, um dos atributos do conjunto de teste (no caso *Relação*) e o algoritmo J48 foi selecionado para ser aplicado segundo a abordagem *cross-validation*.

Por fim, na Seção 5, é possível consultar o classificador gerado, além de se analisar as medidas de avaliação e a matriz de confusão, também geradas automaticamente. Tais conceitos são detalhados a seguir.

De acordo com Hirschman e Mani (2003), utilizam-se extensivamente as medidas Precisão, Cobertura e Medida-F para avaliar o desempenho dos classificadores.

A *Precisão* é a quantidade de instâncias corretamente classificadas em relação à quantidade total de instâncias. Assim, é possível verificar a eficácia do classificador em retornar a informação tida como relevante. Em (26) demonstra-se o cálculo da Precisão.

(26)

$$Precisão = \frac{TP}{TP + FP}$$

Em (26), demonstra-se o cálculo da Precisão, que se baseia na quantidade de casos *true positives* (TP) em razão da quantidade somada de casos verdadeiros e *false positives* (FP).

A *Cobertura* é a quantidade de casos corretamente detectados em relação à quantidade que deveria ser detectada. Assim, é possível avaliar a quantidade de instâncias são cobertas pelo classificador. Em (27) demonstra-se o cálculo da Cobertura.

(27)

$$Cobertura = \frac{TP}{TP + FN}$$

Em (27), demonstra-se o cálculo da Cobertura, que se baseia na quantidade de casos TP em razão da quantidade somada de casos verdadeiros e FN.

Por fim, para avaliar a eficácia de um classificador, utiliza-se a Medida-F, que é a média ponderada das medidas anteriores, como demonstrado em (28).

(28)

$$MedidaF = \frac{2\ (Precisão.Cobertura)}{(Precisão + Cobertura)}$$

Em 28, demonstra-se o cálculo da Medida-F, que se baseia na dupla quantidade multiplicativa de Precisão e Cobertura em razão da quantidade somada de Precisão e Cobertura.

Outra maneira de avaliar um classificador é observar a Matriz de Confusão gerada pelo Weka durante o processo de desenvolvimento do classificador. Na matriz, é possível analisar resultados de maneira inequívoca. Dado um classificador a quantidade de relações representará a quantidade de entradas na tabela. Em uma matriz com duas relações, *X* e *Y*, tem-se uma tabela de duas entradas: a de classe desejada e a de classe preterida. As células, então, são preenchidas com a quantidade de instâncias correspondentes ao cruzamento das entradas, como demonstrado na Tabela 10.

X	Y	
10	2	X
2	10	Y

Tabela 10: Exemplo de Matriz de Confusão.
Fonte: Elaborado pelo autor.

Na Tabela 10, exemplifica-se uma matriz de confusão. A matriz é resultado de uma análise automática sobre 12 instâncias, que foram classificadas entre as relações *X* e *Y*. Ao realizar a categorização das instâncias, quando o classificador admite a relação *X*, por exemplo, duas instâncias têm o rótulo da relação *Y*, o que acontece inversamente quando o classificador identifica a relação *Y*. Com base nesse tipo de análise, é possível perceber a proximidade ou distanciamento entre as relações, e investigar qual o motivo que levou a esse resultado.

Na próxima subseção, apresenta-se a tarefa de seleção de atributos, também disponível no Weka.

5.3. Seleção dos algoritmos para o Aprendizado de Máquina

Com o objetivo de investigar a relevância e dispositivos de sinalização, selecionaram-se os algoritmos supervisionados utilizados por Maziero (2012) e Souza *et al.* (2015; 2018), a saber: (i) PART (WITTEN, FRANK, 1998), (ii) J48 (QUILAN, 1993), (iii) OneR (HOLTE, 1993).

Os algoritmos do paradigma supervisionado evidenciam o conhecimento implícito a partir de exemplos previamente classificados, gerando classificadores (p.ex.: conjunto de regras) que relacionam os atributos (e seus valores) às classes. Neste trabalho, esses algoritmos partiram de um *corpus* de pares de sentenças anotados com as relações CST de complementaridade e atributos (isto é, características das instâncias que podem ser relevantes para o aprendizado das relações) e buscaram identificar padrões/regularidades que evidenciam a correlação entre as classes e os atributos, originando os classificadores.

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 (ii) os atributos/valores das classes. Tal conjunto de dados é ilustrado pela Tabela 10. 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, também do paradigma simbólico, constrói classificadores em formato de árvores de decisão. A abordagem utilizada durante a construção da árvore é *top-down*. A partir de um conjunto de atributos, elenca-se o mais significativo, o qual é tido como o nó inicial da árvore; consequentemente, os nós subsequentes serão os menos significativos em relação aos anteriores.

Por fim, o algoritmo One-R 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.

5.4. Criação dos classificadores

Nesta seção, serão demonstrados os classificadores desenvolvidos a partir dos algoritmos PART, J48 e One-R. Tais classificadores passaram por 2 testes, a saber:

(i) geração de classificadores com base em todos os atributos da tipologia de sinalizadores, utilizando-se a estratégia *cross-fold validation* (10 *folds*);

(ii) geração de classificadores com base em sinalizadores computacionalmente processáveis, utilizando-se a estratégia *cross-fold validation* (10 *folds*);

A técnica utilizada para a geração dos classificadores foi a *cross-fold validation.* Nessa técnica, as instâncias submetidas aos algoritmos são aleatoriamente divididas partições (*folds*) mutuamente exclusivas de tamanho proporcional e aproximadamente iguais. Diante da quantidade de partições, são realizados experimentos em quantidade semelhante (p.ex. se são feitos 10 *folds*, ter-se-ão 10 experimentos) em que, em cada experimento, uma partição é escolhida para o teste e as outras restantes são escolhidas para treinamento. Essa técnica é aconselhada quando o conjunto de instâncias analisado é relativamente pequeno, impossibilitando o treinamento e teste em conjuntos distintos.

Ademais, realizaram-se testes para a geração dos classificadores que visasse a menor quantidade possível de regras frente ao maior número de acertos. Na interface gráfica do Weka adaptaram-se as propriedades dos algoritmos para cumprir tal objetivo. Dessa maneira, os parâmetros *confidence fator* (responsável por aumentar a precisão do classificador) e *minNumObj* (responsável pela quantidade máxima de regras) parametrizados para 0.5 e 3, respectivamente.

Na próxima subseção discrimina-se a tarefa de seleção de atributos realizada durante o processo de desenvolvimento dos classificadores.

5.4.1. Seleção de atributos

A fase de seleção dos atributos, em AM, é importante para analisar o *hall* de atributos que, possivelmente, possam ser irrelevantes para o processo de elaboração dos classificadores.

De maneira geral, os algoritmos tendem a ter melhor desempenho quanto à precisão quando lidam com menos atributos, elevando o grau de aprendizado ao que se refere à classificação das instâncias. Outro motivo para que haja a seleção de atributos é o fato de que, em PLN, a caracterização linguística de grandes *corpora* é cara. Assim, partindo-se da seleção de atributos levantou-se um pequeno conjunto de características que, a partir delas, possam ser atribuídas ao *corpus*.

Nesta pesquisa, a fase de seleção de atributos foi realizada manualmente e automaticamente. Na fase manual, percebeu-se que algumas características linguísticas não eram recorrentes nas instâncias anotadas com as relações CST de complementaridade. Tais

atributos, durante a elaboração da tipologia, foram removidos para que, posteriormente, fossem aplicados no *corpus* de treinamento/teste. Na fase automática, por meio do algoritmo *InfoGainAttributeEval*, levantaram-se os atributos mais importantes por meio de um *ranking* criado pelo próprio algoritmo.

Entretanto, a fase automática foi realizada, mas não foi aplicada a este estudo, uma vez que os algoritmos escolhidos já fazem a seleção de atributos para desenvolver os classificadores. Dessa maneira, a tarefa de seleção automática de atributos deu suporte às análises realizadas sobre os classificadores.

Na Tabela 11 e 12 tem-se os *rankings* de atributos: na primeira tabela, o ranking utilizando todos os atributos da tipologia construída neste estudo; na segunda, o ranking utilizando somente os atributos computacionalmente tratáveis.

RAKING	DESEMPENHO	ATRIBUTO
1	36,59%	EXPRESSÃO TEMPORAL
2	32,69%	SIMILARIDADE DE EVENTOS
3	31,67%	FREQUÊNCIA
4	31,67%	FATO PONTUAL
5	27,70%	EVENTO POSTERIOR
6	20,30%	EXPRESSÃO SUPERLATIVA
7	17,10%	TEMPO VERBAL
8	16,17%	TEMA-REMA
9	10,85%	DETALHE DO EVENTO
10	10,28%	DISCURSO REPORTADO
11	9,98%	VERBOS DE ELOCUÇÃO
12	9,97%	PREPOSIÇÃO
13	8,19%	CAMPO SEMÂNTICO
14	8,01%	PROLONGAMENTO DO EVENTO
15	8,01%	INFORMAÇÃO ADICIONAL
16	6,30%	ADJUNTO ADVERBIAL
17	5,01%	PROJEÇÃO FUTURA
18	5,01%	ORAÇÃO ADITIVA
19	4,71%	CONHECIMENTO DE MUNDO
20	4,28%	HIPONÍMIA
21	2,94%	PARTE-TODO
22	2,83%	EXPRESSÃO NOMINAL
23	2,83%	ORAÇÃO REDUZIDA
24	2,80%	ORAÇÃO OBJETIVA DIRETA
25	2,42%	ANÁFORA NOMINAL
26	2,16%	LEITURA DO CLUSTER
27	2,11%	PREVISÃO DO EVENTO
28	1,48%	NUMERAL

29	1,37%	CAUSA-EFEITO
30	1,02%	ANÁFORA ASSOCIATIVA
31	0,38%	SEMÂNTICA LEXICAL
32	0,36%	ORAÇÃO EXPLICATIVA

Tabela 11: *Ranking* de seleção de atributos.
Fonte: Elaborado pelo autor.

RAKING	DESEMPENHO	ATRIBUTO
1	37%	EXPRESSÃO TEMPORAL
2	20%	EXPRESSÃO SUPERLATIVA
3	17%	TEMPO VERBAL
4	10%	DISCURSO REPORTADO
5	10%	VERBOS DE ELOCUÇÃO
6	10%	PREPOSIÇÃO
7	6%	ADJUNTO ADVERBIAL
8	5%	ORAÇÃO ADITIVA
9	4%	HIPONÍMIA
10	3%	PARTE-TODO
11	3%	ORAÇÃO REDUZIDA
12	3%	EXPRESSÃO NOMINAL
13	3%	ORAÇÃO OBJETIVA DIRETA
14	1%	NUMERAL
15	1%	CAUSA-EFEITO
16	0%	ORAÇÃO EXPLICATIVA

Tabela 12: Ranking de seleção de atributos computacionalmente tratáveis.
Fonte: Elaborado pelo autor.

5.4.2. Classificadores gerados com base no conjunto total dos sinais

Com base na tipologia proposta neste trabalho (ver pág. 72), construíram-se classificadores com base em todos os dispositivos presentes na tipologia. Dentre eles, alguns não são computacionalmente tratáveis, ou não estão na superfície textual.

a) One-R

No Quadro 16, tem-se o classificador obtido.

```
EXPRESSÃO TEMPORAL
        NÃO  → ELABORATION
        SIM  → HISTORICAL BACKGROUND
```

Quadro 16: Classificador gerado a partir do algoritmo automático One-R.
Fonte: Elaborado pelo autor.

No Quadro 16, tem-se o classificador obtido pelo algoritmo One-R. As regras do classificador basearam-se apenas no dispositivo e*xpressão temporal*. A primeira regra do classificador indica, por exemplo, que a relação será *Elaboration* se não houver esse sinalizador no par de sentença. Caso contrário, a relação será *Historical background*. A taxa de acerto desse classificador foi de 58,3%, ou seja, 133 pares de sentença foram classificadas corretamente, dos 228 que compunham o *subcorpus* de treinamento/teste. Na Tabela 13, apresentam-se os resultados das medidas de avaliação do classificador.

RELAÇÃO CST	PRECISÃO	COBERTURA	MEDIDA-F
ELABORATION	0.53	1.0	0.69
FOLLOW-UP	0	?	?
HISTORICAL BACKGROUND	0.66	0.75	0.70

Tabela 13: Medidas de avaliação do classificador gerado pelo algoritmo One-R.
Fonte: Elaborado pelo autor.

Com base na Tabela 13, conclui-se que sempre que o classificador se deparou com um par de sentenças anotado com a relação *Historical Background* ele alcançou precisão de 66% e obteve cobertura de 75% das instâncias. Sobre a relação *Elaboration*, é importante destacar que o classificador tem uma precisão de 53%, mas cobriu 100% das instâncias. Não há dados sobre a relação *Follow-up* porque o classificador elegeu um dispositivo de sinalização cujos valores são binários (a saber, *Sim* ou *Não*), e previu que tal sinalizador (no caso, a ET) seria mais relevante na distinção entre *Elaboration* e *Historical Background*.

Por fim, na Tabela 14 apresentam-se os dados da Matriz de confusão do classificador gerado.

ELABORATION	***FOLLOW-UP***	***HISTORICAL BACKGROUND***	
76	0	0	***ELABORATION***
47	0	29	***FOLLOW-UP***
19	0	57	***HISTORICAL BACKGROUND***

Tabela 14: Matriz de confusão do classificador One-R com todos os atributos da tipologia.
Fonte: Elaborado pelo autor.

Os resultados da Tabela 12 indicam que o classificador possui limitações quanto à distinção das relações CST. Utilizando a regra proposta em instâncias anotadas com a relação *Elaboration*, por exemplo, o classificador acertou todas as 65 instâncias. Por outro lado, valendo-se dessa mesma regra, identificou 17 pares anotados com *Historical Background* como sendo *Elaboration*.

b) PART

Baseado no modelo de regras lógicas "*se 'x', então 'y', senão 'z'*", o algoritmo PART cria um classificador cujas regras são dependentes da condição de verdade umas das outras. Isso quer dizer que para que a segunda regra seja verdadeira, a primeira deve ser falsa; para que a terceira seja verdadeira, a primeira e a segunda devem ser falsas, e assim por diante. No Quadro 17, tem-se o classificador gerado pelo algoritmo PART.

1. **SE** SIMILARIDADE DO EVENTO = NÃO **E** FREQUÊNCIA = NÃO **E** FATO PONTUAL = NÃO **E** EVENTO POSTERIOR = NÃO **E** EXPRESSÃO TEMPORAL = NÃO **E** PROLONGAMENTO DO EVENTO = NÃO **E** PROJEÇÃO FUTURA = NÃO **E** TEMPO VERBAL = NÃO, **ENTÃO** *ELABORATION*

2. **SENÃO**, EVENTO POSTERIOR = NÃO **E** TEMPO VERBAL = NÃO **E** CAMPO SEMÂNTICO = NÃO **E** DETALHE DO EVENTO = NÃO **E** FREQUÊNCIA = SIM, **ENTÃO** *HISTORICAL BACKGROUND*

3. **SENÃO**, FATO PONTUAL = NÃO, **ENTÃO** *FOLLOW-UP*
4. **CASO CONTRÁRIO**, *HISTORICAL BACKGROUND*

Quadro 17: Classificador gerado pelo algoritmo PART.

Fonte: Elaborado pelo autor.

Com base no Quadro 17, a Regra 1 determina que a relação CST é *Elaboration* caso os sinalizadores similaridade do evento, fato pontual, evento posterior, expressão temporal, prolongamento do evento, projeção futura e tempo verbal tenham valor *Não*. Caso a Regra 1 não seja suficiente a todas às instâncias, aciona-se a Regra 2 e, caso ela continue não sendo suficiente, o classificador aplicará as demais regras. Ao final, se nenhuma das três regras for verdadeira, a relação *default* será *Historical background*.

Em relação ao classificador gerado pelo algoritmo One-R, houve aumento significativo da performance do classificador, já que obteve 93,4% de acerto, isto é, 213 instâncias corretamente classificadas. Na Tabela 15, tem-se o resultado das medidas de avaliação do algoritmo.

RELAÇÃO CST	PRECISÃO	COBERTURA	MEDIDA-F
ELABORATION	0.89	0.96	0.92
FOLLOW-UP	0.95	0.84	0.89
HISTORICAL BACKGROUND	0.96	1.0	0.98

Tabela 15: Medidas de avaliação do classificador gerado pelo algoritmo PART.
Fonte: Elaborado pelo autor.

Da Tabela 15, observa-se que o classificador apresenta medidas-f superiores para as relações *Historical Background* e *Elaboration* (92% e 98%, respectivamente); já quanto à *Follow-up*, o classificador PART alcançou resultado inferior, porém expressivo (85%). Salienta-se que o classificador gerado pelo PART supera o estado-da-arte (SOUZA, 2015), que é de 75% de medida-f para a relação *Historical Background* e de 66% para as outras relações, utilizando o mesmo algoritmo.

Na Tabela 16, apresentam-se os resultados da Matriz de confusão gerada a partir do classificador descrito.

ELABORATION	***FOLLOW-UP***	***HISTORICAL BACKGROUND***	
73	3	0	***ELABORATION***
9	64	3	***FOLLOW-UP***
0		76	***HISTORICAL BACKGROUND***

Tabela 16: Matriz de confusão do classificador PART com todos os atributos da tipologia.
Fonte: Elaborado pelo autor.

A partir dos resultados da Tabela 16, infere-se que, como o classificador conta com mais dispositivos de sinalização para compor suas regras, ele indica com mais precisão as relações CST. Quando o classificador se depara com instâncias anotadas com a relação *Historical background*, ele não apresenta dúvidas de classificação entre essas relações de complementaridade, o que é indicado por zero número de casos confundidos. Diante de instâncias anotadas com a relação *Elaboration*, não há equívocos de classificação com a relação *Historical background*, porém confunde-se discretamente com a relação Follow-up (3 instâncias, apenas). Por fim, quanto à relação *Follow-up* pode-se considerar que o desempenho do classificador gerado é expressivo, já que, frente à quantidade de instâncias, há poucos casos de equívocos com as outras relações CST de complementaridade.

c) J48

O algoritmo J48 constrói classificadores de maneira bastante semelhante ao PART, já que suas regras são dependentes entre si. Entretanto, a o J48 segue o padrão de regras por decisão, em que se constroem caminhos (estatisticamente) possíveis a partir do atributo mais relevante julgado pelo algoritmo. No Quadro 18, tem-se o classificador gerado.

```
FREQUÊNCIA = NÃO
|  FATO PONTUAL = NÃO
|  |  EVENTO POSTERIOR = NÃO
|  |  |  EXPRESSÃO TEMPORAL = NÃO
|  |  |  |  PROLONGAMENTO DO EVENTO = NÃO
|  |  |  |  |  PROJEÇÃO FUTURA = NÃO
|  |  |  |  |  |  PREVISÃO DO EVENTO = NÃO
|  |  |  |  |  |  |  TEMPO VERBAL = NÃO
|  |  |  |  |  |  |  |  ANÁFORA ASSOCIATIVA = NÃO, ENTÃO ELABORATION
|  |  |  |  |  |  |  |  ANÁFORA ASSOCIATIVA = SIM
|  |  |  |  |  |  |  |  |  DISCURSO REPORTADO = SIM, ENTÃO FOLLOW-UP
|  |  |  |  |  |  |  |  |  DISCURSO REPORTADO = NÃO, ENTÃO ELABORATION
|  |  |  |  |  |  |  TEMPO VERBAL = SIM
|  |  |  |  |  |  |  |  DETALHE DO EVENTO = NÃO, ENTÃO ELABORATION
|  |  |  |  |  |  |  |  DETALHE DO EVENTO = SIM, ENTÃO FOLLOW-UP
|  |  |  |  |  |  PREVISÃO DO EVENTO = SIM, ENTÃO FOLLOW-UP
|  |  |  |  |  PROJEÇÃO FUTURA = SIM, ENTÃO FOLLOW-UP
|  |  |  |  PROLONGAMENTO DO EVENTO = SIM, ENTÃO FOLLOW-UP
|  |  |  EXPRESSÃO TEMPORAL = SIM, FOLLOW-UP
|  |  EVENTO POSTERIOR = SIM, ENTÃO FOLLOW-UP
|  FATO PONTUAL = SIM , ENTÃO HISTORICAL BACKGROUND
FREQUÊNCIA = SIM, ENTÃO HISTORICAL BACKGROUND
```

Quadro 18: Classificador gerado pelo algoritmo J48.
Fonte: Elaborado pelo autor.

No Quadro 18, o conjunto de regras de decisão aponta que o sinalizador mais genérico e significativo é a *frequência*. Caso tal dispositivo de sinalização tenha o valor "SIM", a relação é *Historical background*. No entanto, se o mesmo dispositivo obtiver valor "NÃO", é necessário considerar os dispositivos *fato pontual*, evento posterior, expressão temporal, prolongamento do evento, projeção futura, previsão do evento, tempo verbal e anáfora associativa com valor "NÃO" para ser a relação *Elaboration*. Caso tais regras não classifiquem todas as instâncias, serão consideradas outras regras compostas ao longo da árvore de decisão.

Ainda com base nesse classificador, nota-se que, como dito anteriormente, os sinalizadores exclusivos de cada relação CST de complementaridade cumprem papel de extrema importância na geração dos classificadores, ocupando os três lugares mais genéricos na árvore, no caso, os sinalizadores *evento posterior* (característico e exclusivo de *Follow-up*), *frequência* e *fato pontual* (característicos e exclusivos de *Historical background*). Salienta-se que tais sinalizadores advêm da análise manual e da proposta dos sinalizadores da categoria pragmático, os quais visam a suprir a lacuna da adaptação do conjunto de relações do Inglês para o PB, apontado anteriormente. Como resultado, o classificador identifica corretamente 217 instâncias, ou seja, 95,17% do *subcorpus* de treinamento/teste.

Na Tabela 17, têm-se as medidas de avaliação do classificador do Quadro 18.

RELAÇÃO CST	PRECISÃO	COBERTURA	MEDIDA-F
ELABORATION	0.82	0.98	0.93
FOLLOW-UP	0.98	0.86	0.92
HISTORICAL BACKGROUND	1.0	1.0	1.0

Tabela 17: Medidas de avaliação do classificador pelo algoritmo J48.
Fonte: Elaborado pelo autor.

Depreende-se da Tabela 17 que, apesar do classificador apresentar valores discretamente baixos quanto à identificação da relação *Follow-up*, a medida-f de todas as três relações ficam próximas ou são superiores a 92%. Tendo em vista que o classificador utiliza sinalizadores exclusivos de, ao menos, duas relações de complementaridade, infere-se que sua robustez e acerto sejam superiores. Isso reflete-se na baixa confusão que o classificador faz ao identificar as relações das instâncias, uma vez que apenas as relações *Follow-up* e *Elaboration* são discretamente confundidas entre si, como é demonstrado na Tabela 18.

ELABORATION	*FOLLOW-UP*	*HISTORICAL BACKGROUND*	
75	1	0	*ELABORATION*
10	66	3	*FOLLOW-UP*
0	0	76	*HISTORICAL BACKGROUND*

Tabela 18: Matriz de confusão do classificador J48 com todos os atributos da tipologia.
Fonte: Elaborado pelo autor.

Na próxima subseção, apresentam-se os classificadores construídos a partir dos dispositivos de sinalização que são computacionalmente tratáveis.

5.4.3. Classificadores gerados com base nos dispositivos computacionalmente tratáveis

Nessa subseção, apresentam-se os classificadores gerados com base nos dispositivos linguísticos que, atualmente, podem ser automaticamente tratáveis no cenário do PLN, destacando os repositórios e/ou ferramentas necessários a esse processamento. Destaca-se que, embora os testes que serão apresentados, a seguir, estejam lidando com os atributos computacionalmente tratáveis, o desempenho de um método de detecção das relações com base nesses atributos dependerá, sobretudo, do desempenho das ferramentas e da robustez dos recursos de PLN empregados. Nesse sentido, os resultados que serão demonstrados são apenas indícios de como poderia ser a detecção das relações com base nesse conjunto de atributos. Assim, retomando o que aponta Taboada e Das (2013) sobre a descrição de relações de sentido, ressalta-se que nem sempre as pistas (ou sinalizadores) podem ser reconhecidas

por sistemas automáticos, mas sempre são reconhecidas por humanos. Isso significa admitir que as descrições linguísticas nem sempre são modeladas computacionalmente em sua totalidade.

No Quadro 19, demonstram-se os sinais (genéricos e específicos) cujo tratamento computacional pode ser subsidiado por repositórios de conhecimento e/ou ferramentas de PLN.

Tipologia - PLN			Recursos/Ferramentas de PLN
Categoria	**Sinal genérico**	**Sinal específico**	
Morfológico	**Classe de palavras**	Numeral	Análise sintática (p.ex. Palavras -bick, 2000)
		Expressão Nominal	
		Preposição	
	Temporal	Tempo verbal	
	Verbos de elocução	Verbos de elocução	Léxico de Verbos de elocução (Costa; freitas, 2017)
Sintático	**Período simples**	Adjunto adverbial	Análise sintática (p.ex. Palavras -bick, 2000)
		Expressão superlativa	
	Período composto	Discurso reportado	
		Oração aditiva	
		Oração explicativa	
		Oração objetiva direta	
		Oração reduzida	
Semântico	**Relações semânticas**	Causa-Efeito	Base de dados (p.ex. WordNet – Fellbaum, 1998)
		Hiponímia	
		Parte-Todo	
	Temporal	Expressão temporal	Anotação de Expressões Temporais (Menezes-Filho; Pardo, 2008)

Quadro 19: Lista de dispositivos de sinalização da complementaridade computacionalmente tratáveis.
Fonte: Elaborado pelo autor.

Nesse novo cenário de treinamento, apenas 16 dispositivos de sinalização foram considerados aplicáveis computacionalmente. Além disso, é possível prever que o alcance dos classificadores que serão gerados com base nesse conjunto de sinalizadores será comprometido, uma vez que os dispositivos que foram identificados como mais significativos nos testes anteriores foram removidos.

Salienta-se que os classificadores a seguir foram construídos com base em uma caracterização da complementaridade de maneira binária, sendo assinalados pelos valores S (para "sim") e N (para "não"). Assim, dado par de sentenças (S1,S2), S2 indicasse que o sinalizador "Oração Aditiva", por exemplo, fosse a característica que caracteriza a relação *Follow-up*, tal sinalizador receberia o valor "S"; caso contrário, "N".

a) One-R

Após a remoção dos dispositivos não processáveis computacionalmente, o algoritmo elegeu o sinalizador Tempo Verbal como o mais significativo. No Quadro 20, tem-se o classificador gerado.

```
EXPRESSÃO TEMPORAL
    NÃO  → ELABORATION
    SIM  → HISTORICAL BACKGROUND
```

Quadro 20: Classificador gerado pelo automático One-R com base em dispositivos processáveis.
Fonte: Elaborado pelo autor.

De acordo com o classificador, a relação CST será *Elaboration* caso não haja Expressão Temporal no par de sentença, ao passo que será *Historical background* se tal dispositivo ocorrer. Já que a caracterização foi binária, a relação *Follow-up* não foi contemplada nessa regra, pois, possivelmente, tal dispositivo de sinalização não foi considerado como relevante nos pares de sentenças que compõem o *subcorpus* de treinamento/teste. Como resultado, classificaram-se corretamente 133 instâncias, ou seja, 58,33% dos pares de sentenças.

Na Tabela 19, têm-se os resultados das medidas de avaliação do classificador do Quadro 20.

RELAÇÃO CST	PRECISÃO	COBERTURA	MEDIDA-F
ELABORATION	0.53	1.0	0.69
FOLLOW-UP	?	0	?
HISTORICAL BACKGROUND	0.66	0.75	0.70

Tabela 19: Avaliação do classificador gerado por One-R com base em dispositivos processáveis.
Fonte: Elaborado pelo autor.

Da Tabela 19, nota-se que o classificador possui uma cobertura maior para a relação *Elaboration* (isto é, 100%). Porém, o One-R apresenta uma precisão maior para a relação *Historical background* (isto é, 66%). Provavelmente, esse classificador não seja relevante para determinar qual seja a relação CST de complementaridade, mas o tipo de complemento proposto por Maziero *et al.* (2010), (no caso, temporal e atemporal), uma vez que a regra se baseia em um dispositivo temporal em que se determina sua presença ou ausência.

Além disso, ressalta-se que, mesmo com o comprometimento das medidas avaliativas nesse cenário, o classificador proposto aqui é equivalente ao de Souza (2015). Na ocasião, baseando-se no atributo *expressão temporal*, o autor utilizou atributos processáveis e não-processáveis computacionalmente e obteve 68% de medida-f para a relação *Elaboration*.

Na Tabela 20, apresentam-se os resultados da matriz de confusão do classificador gerado neste cenário.

ELABORATION	***FOLLOW-UP***	***HISTORICAL BACKGROUND***	
76	0	0	***ELABORATION***
47	0	29	***FOLLOW-UP***
19	0	57	***HISTORICAL BACKGROUND***

Tabela 20: Matriz de confusão do classificador One-R com todos os atributos da tipologia.
Fonte: Elaborado pelo autor.

Ressalta-se que a matriz de confusão desse classificador é semelhante a apresentada na subseção anterior, já que o algoritmo One-R utilizou o mesmo sinalizador (no caso, Expressão temporal) para criar o conjunto de regras proposto.

b) Part

No Quadro 21, tem-se o conjunto de regras criado pelo algoritmo PART.

1. **SE** EXPRESSÃO SUPERLATIVA = NÃO **E** EXPRESSÃO TEMPORAL = SIM **E** VERBOS DE ELOCUÇÃO = NÃO **E** TEMP_TEMPVERB = N **E** ADJUNTO ADVERBIAL = NÃO, **ENTÃO** *HISTORICAL BACKGROUND*

2. **SENÃO** EXPRESSÃO SUPERLATIVA = NÃO **E** EXPRESSÃO TEMPORAL = NÃO **E** TEMPO VERBAL = NÃO **E** VERBOS DE ELOCUÇÃO = NÃO **E** NUMERAL = NÃO **E** HIPONÍMIA = NÃO **E** ORAÇÃO OBJETIVA DIRETA = NÃO **E** ADJUNTO ADVERBIAL = NÃO **E** ORAÇÃO EXPLICATIVA = NÃO, **ENTÃO** *ELABORATION*

3. **SENÃO** EXPRESSÃO SUPERLATIVA = SIM, **ENTÃO** *HISTORICAL BACKGROUND*

4. **SENÃO** HIPONÍMIA = SIM, **ENTÃO** *ELABORATION*

5. **SENÃO** ORAÇÃO EXPLICATIVA = NÃO **E** ORAÇÃO OBJETIVA DIRETA = SIM **E** DISCURSO REPORTADO = SIM, **ENTÃO** *ELABORATION*

6. **SENÃO** ORAÇÃO REDUZIDA = NÃO **E** ORAÇÃO EXPLICATIVA = NÃO, **ENTÃO** *FOLLOW-UP*

7. **SENÃO** ADJUNTO ADVERBIAL = NÃO **E** ORAÇÃO REDUZIDA = NÃO, **ENTÃO** *HISTORICAL BACKGROUND*

8. **SENÃO** ADJUNTO ADVERBIAL = NÃO, **ENTÃO** *ELABORATION*

9. **CASO CONTRÁRIO**, *FOLLOW-UP*

Quadro 21: Classificador gerado pelo automático PART com dispositivos processáveis computacionalmente.
Fonte: Elaborado pelo autor.

De acordo com o Quadro 21, tem-se 9 regras para classificar as relações CST de complementaridade. Além da diminuição de instâncias corretamente classificadas (neste caso, 71,2%, o que representa 164 instâncias), ressalta-se o fato de a relação *default* ter passado a

ser *Follow-up*, e não mais *Historical background* quando o cenário de treinamento conta com todos os sinalizadores.

Na Tabela 21 têm-se os resultados das medidas de avaliação desse classificador.

RELAÇÃO CST	PRECISÃO	COBERTURA	MEDIDA-F
ELABORATION	0.64	0.73	0.68
FOLLOW-UP	0.72	0.59	0.65
HISTORICAL BACKGROUND	0.79	0.82	0.81

Tabela 21: Avaliação do classificador gerado por PART com dispositivos processáveis computacionalmente..
Fonte: Elaborado pelo autor.

Tendo como ponto de partida à discussão a medida-f, os resultados indicados na Tabela 21 aproximam-se dos resultados da mesma medida para o classificador One-R, o qual se baseia em apenas um sinalizador processável no PLN. Além disso, é necessário considerar que, caso o processo de identificação desses sinalizadores processáveis for automatizado, é possível que as medidas de avaliação demonstradas aqui diminuam, refletindo-se, inclusive, nos equívocos que o classificador pode apresentar ao atribuir as relações às instâncias.

Assim, na Tabela 22, demonstra-se a matriz de confusão desse classificador.

ELABORATION	*FOLLOW-UP*	*HISTORICAL BACKGROUND*	
56	13	7	*ELABORATION*
22	45	9	*FOLLOW-UP*
9	4	63	*HISTORICAL BACKGROUND*

Tabela 22: Matriz de confusão do classificador PART com dispositivos processáveis computacionalmente.
Fonte: Elaborado pelo autor.

Com base na Tabela 22, o classificador apresenta um nível maior de equívocos. Tal fato leva à inferência de que a ausência de atributos semânticos e sintáticos interferem significativamente no processo de identificação das relações CST, além de demonstrar que apenas sinalizadores que manifestem a informação temporal são úteis, mas não suficientes para esta tarefa.

c) J48

No Quadro 22 tem-se o classificador gerado pelo algoritmo J48.

```
EXPRESSÃO SUPERLATIVA = NÃO
|  EXPRESSÃO TEMPORAL =NÃO
|  |  TEMPO VERBAL = NÃO
|  |  |  VERBOS DE ELOCUÇÃO = NÃO
|  |  |  |  NUMERAL = NÃO
|  |  |  |  |  ORAÇÃO OBJETIVA DIRETA = NÃO
|  |  |  |  |  |  ADJUNTO ADVERBIAL = NÃO
|  |  |  |  |  |  |  ORAÇÃO EXPLICATIVA = NÃO, ENTÃO ELABORATION
|  |  |  |  |  |  |  ORAÇÃO EXPLICATIVA = SIM, ENTÃO HISTORICAL BACKGROUND
|  |  |  |  |  |  ADJUNTO ADVERBIAL = SIM, ENTÃO FOLLOW-UP
|  |  |  |  |  ORAÇÃO OBJETIVA DIRETA = SIM, ENTÃO ELABORATION
|  |  |  |  NUMERAL = SIM, ENTÃO FOLLOW-UP
|  |  |  VERBOS DE ELOCUÇÃO = SIM, ENTÃO FOLLOW-UP
|  |  TEMPO VERBAL = SIM, FOLLOW-UP
|  EXPRESSÃO TEMPORAL = SIM
|  |  VERBOS DE ELOCUÇÃO = NÃO
|  |  |  TEMPO VERBAL = NÃO
|  |  |  |  ADJUNTO ADVERBIAL = NÃO, ENTÃO HISTORICAL BACKGROUND
|  |  |  |  ADJUNTO ADVERBIAL = SIM, ENTÃO FOLLOW-UP
|  |  |  TEMPO VERBAL = SIM, ENTÃO FOLLOW-UP
|  |  VERBOS DE ELOCUÇÃO = SIM, ENTÃO FOLLOW-UP
EXPRESSÃO SUPERLATIVA = SIM, ENTÃO HISTORICAL BACKGROUND
```

Quadro 22: Classificador gerado pelo algoritmo J48 com dispositivos processáveis computacionalmente.
Fonte: Elaborado pelo autor.

Depreende-se que a partir do conjunto de regras gerados nesse cenário pelo algoritmo J48 que para o classificador é muito mais custoso indicar corretamente as instâncias anotadas com as relações CST de complementaridade, tendo em vista o tamanho da árvore de decisão, que resultou em 172 instâncias corretamente classificadas (o que representa 75,4% do *subcorpus* de treinamento/teste).

Nas Tabelas 23 e 24 têm-se os resultados das medidas de classificação e a matriz de confusão, respectivamente, do classificador gerado pelo algoritmo J48 no cenário de apenas utilizar sinalizadores processáveis no PLN. Como previsto anteriormente, os resultados ainda são expressivos, apesar de decaírem devido a remoção massiva dos dispositivos de sinalização de natureza semântica e pragmática.

RELAÇÃO CST	PRECISÃO	COBERTURA	MEDIDA-F
ELABORATION	0.68	0.75	0.71
FOLLOW-UP	0.75	0.69	0.72
HISTORICAL BACKGROUND	0.82	0.81	0.82

Tabela 23: Medidas de avaliação do classificador gerado pelo algoritmo J48 com dispositivos processáveis computacionalmente.
Fonte: Elaborado pelo autor.

ELABORATION	***FOLLOW-UP***	***HISTORICAL BACKGROUND***	
57	13	6	***ELABORATION***
16	53	7	***FOLLOW-UP***
10	4	62	***HISTORICAL BACKGROUND***

Tabela 24: Matriz de confusão do classificador gerado pelo algoritmo J48 com dispositivos processáveis computacionalmente.
Fonte: Elaborado pelo autor.

No próximo capítulo tecem-se considerações finais acerca do trabalho ora descrito, além de suas contribuições nas áreas de Linguística e PLN.

Capítulo 6

CONSIDERAÇÕES FINAIS

Neste capítulo teço considerações finais sobre esta pesquisa, destaco as limitações ao seu desenvolvimento e à área, além de pontuar contribuições pertinentes à Linguística e ao PLN.

Neste trabalho, aprofundou-se a descrição sobre o fenômeno da complementaridade, que ocorre em conjuntos de textos jornalísticos que abordam um mesmo assunto.

Especificamente, estendeu-se a descrição já realizada por Souza (2015) que, ao se basear apenas em atributos que manifestavam informações temporais nas sentenças de um par, já havia obtido resultados bastantes satisfatórios. Assim, tomando o trabalho de Souza (2015) como ponto de partida, os desafios deste trabalho foram: (i) linguisticamente, propor atributos que apresentem percepções mais profundas sobre a linguagem, o que advém da maturidade e reflexão sobre o objeto e (ii) computacionalmente, gerar classificadores que superem o desempenho relatado na literatura.

Diferentemente das relações RST que são propositais (ou seja, o autor do texto tem uma intenção expressa que a segunda sentença complemente o sentido/informação da primeira, por exemplo), no modelo CST esse fenômeno é essencialmente "acidental". Os autores dos textos que advêm de fontes distintas não estabelecem "contratos textuais" entre si ao publicarem os textos jornalísticos, determinando que uma sentença do Texto 1 complementará uma outra sentença do Texto 2. Assim, para contornar esse desafio, o fenômeno foi estudado em contextos reais de ocorrência, ou seja, em *corpus*, buscando observar e descrever regularidades linguísticas. Foi necessário, então, dispor de uma quantidade significativa de exemplares para descrever o "comportamento" dos usos da linguagem em um contexto específico (no caso, textos jornalísticos).

Diante do *corpus*, buscou-se pelas pistas linguísticas que supostamente foram utilizadas pelos anotadores quando da rotulação das relações CST, posto que a anotação original não explicita esses indícios. Assim, conseguiu-se compreender os usos linguísticos desse fenômeno em textos jornalísticos: a partir de uma perspectiva mais ampla e menos engessada, como a informação temporal, apenas.

Levantados os dispositivos que sinalizam as sentenças no CSTNews, utilizaram-se algoritmos de AM que auxiliaram na caracterização do fenômeno da complementaridade e das relações CST que o codificam. Nessa perspectiva, os resultados dos classificadores treinados elucidam com mais clareza as informações linguísticas, evidenciando sua robustez de análise e cobertura de instâncias.

Salienta-se que a hipótese traçada inicialmente neste trabalho confirmou-se, em especial, adotar uma perspectiva semântica sobre as relações CST de complementaridade. Tal hipótese foi confirmada, por exemplo, na constituição das regras geradas pelos algoritmos de AM, em que os atributos de natureza semântica e pragmática tem destaque e ampliam a robustez do classificador, uma vez que, removidos, diminuem-na substancialmente.

Acerca das contribuições deste trabalho de doutoramento, salienta-se que elas tocam em duas instâncias naturais a esta linha de pesquisa:

(i) à Linguística Descritiva, ao sistematizar um conjunto amplo de sinalizadores linguísticos da complementaridade que ampliam o conhecimento linguístico que se tinha até então sobre o fenômeno;

(ii) ao PLN, ao fornecer subisídios (sinalizadores linguísticos) para a identificação automática da complementaridade, um dos fenômenos linguísticos mais frequentes nos corpora jornalísticos multidocumento, e cuja identificação pode auxiliar na tarefa de sumarização automática.

Como trabalhos futuros, espera-se aplicar testes estatísticos (como o Teste-T) para verificação das variâncias entre as descrições realizadas por Souza (2015) e as identificadas neste estudo. Além disso, espera-se averiguar qual a recorrência dos atributos descritos neste trabalho em outras relações de redundância e contradição do modelo CST, o que resultaria numa possível reestruturação tipológica das relações.

REFERÊNCIAS

AFANTENOS, S.D., DOURA, I., KAPELLOU, E.; KARKALETSIS,V. Exploiting cross-document relations for multi-document evolving summarization. In Hellenic Conference on Artificial Intelligence. p. 410-419. Springer, Berlin, Heidelberg. 2004.

ANTUNES, R.A.M.R.; PARDO, T.A.S; ALMEIDA, G.M.B. Formação de gentílicos a partir de topônimos: descrição linguística e aprendizado automático (Formation of Demonyms from Toponyms: Linguistic Description and Machine Learning)[In Portuguese]. In: Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology. p. 73-82. 2017.

ALEIXO, P.; PARDO, T.A.S. CSTNews: um corpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory). ICMC-USP. 2008a.

ALEIXO, P. PARDO, T.A.S. CSTTool: Uma ferramenta semi-automática para anotação de córpus pela teoria discursiva multidocumento CST. ICMC-USP. 2008b.

BATISTA, G.E.A.P.A.; PRATI, R.C.; MONARD, M.C. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations, United Stages of America. Junho. v.6, n.1, 20p. 2004.

CARBONELL, J.G.; MICHALSKI, R.S.; MITCHELL, T.M. An overview of machine learning. In: Machine Learning, Volume I. 1983. p. 3-23.

CARDOSO, P.C.F.; MAZIERO, E.G.; JORGE, M.L.C.; SENO, E.M.R.; DI-FELIPPO, A.; RINO, L.H.M.; NUNES, M.G.V.; PARDO, T.A.S. CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the Proceedings of the 3rd RST Brazilian Meeting, p. 88-105. Cuiabá/MT, Brasil. 2011.

CARLETTA, J. Assessing agreement on classification tasks: the kappa statistic. In: Computational linguistics, v. 22, n. 2, p. 249-254. 1996.

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

HATZIVASSILOGLOU, V.; KLAVANS, J.L. ESKIN, E. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In: Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. College Park/Maryland. p.203–212. 1999.

HATZIVASSILOGLOU, V.; KLAVANS J.L.; HOLCOMBE, M. Simfinder: a flexible clustering tool for summarization. In: NAACL AUTOMATIC SUMMARIZATION WORKSHOP. Pittsburgh, PA, USA. Proceedings… p.9. 2001.

HIRSCHMAN, L. MANI ,I. Evaluation. In: MITKOV, R.(org.). The Oxford handbook of computational linguistics. Oxford University Press. 2003.

HOLTE, R.C. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning. v.11, pp 63-90. 1993.

ILARI, R. Articulação Tema-Rema. 2ªed. São Paulo: Ática, 1992.

KRIPPENDORFF, K. Content Analysis: An Introduction to its Methodology. Sage Publications, Beverly Hills, CA. 1980.

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

LAGE, V. Estrutura da notícia. São Paulo: Ática, 1993.

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

MANI, I. Automatic summarization. Vol. 3. John Benjamins Publishing, 2001.

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

MARCUSCHI, L.A. Anáfora sem antecedente explícito. Fala e Escrita em Questão. São Paulo: Humanitas.(Projetos Paralelos–NURC/SP, Núcleo USP, vol. 4). 2000.

MARCUSCHI, L.A. Gêneros textuais: definição e funcionalidade. In: Gêneros textuais e ensino. DIONÍSIO, A.P.; BEZERRA, M.; MACHADO, A.R (orgs). Rio de Janeiro: Lucerna, 2002.

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

MAZIERO, E. G.; PARDO, T. A. S. Multi-Document Discourse Parsing Using Traditional and Hierarchical Machine Learning. In: Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology. 2011.

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

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

MITCHELL, Tom M. Does machine learning really work?. AI magazine, v. 18, n. 3, p. 11, 1997.

MIYABE, Y.; TAKAMURA, H.; OKUMURA, M. Identifying a Cross-Document Relation between Sentences. In: Proceedings of the Third International Joint Conference on Natural Language Processing. v.1. p. 141-148. 2008

PRATI, R.C.; BATISTA, G.E.A.P.A.; MONARD, M.C. Class imbalances versus class overlapping: an analysis of a learning system behavior. In: Mexican international conference on artificial intelligence. Springer, Berlin, Heidelberg, p. 312-321. 2004.

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

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

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

RADEV, D.R.; OTTERBACHER, J. ZHANG, Z. CST Bank: A Corpus for the Study of Cross-document Structural Relationships. In Proceedings of Language Resources and Evaluation Conference (LREC). Lisboa-Portugal. 2004.

SARDINHA, Tony Berber. Lingüística de corpus: histórico e problemática. Delta, v. 16, n. 2, p. 323-367, 2000.

SOUZA, J. W. C. Descrição linguística da complementaridade para a sumarização automática multidocumento. Dissertação (Mestrado em Linguística) – Universidade Federal de São Carlos. p.102. 2015.

SOUZA, J. W. C.; DI FELIPPO, A. Caracterização linguística da complementaridade: subsídios para Sumarização Automática Multidocumento. ALFA: Revista de Linguística (UNESP. ONLINE), 2018.

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. In: Dipper, S.; Zinsmeister, H.; Webber, B. (orgs). Dialogue and Discourse., v.4, n. 2, p. 249-281. 2013.

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

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.; OTTERBACHER, J.; RADEV, D.R. Learning Cross-document Structural Relationsships using Boosting. In: Proceedings of the Twelfth International Conference on Information and knowledge Management. New York. p. 124-130. 2003.

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.

APÊNDICES

Afim de manter preservação ambiental e estimular uma consciência verde, toda esta Tese foi impressa, quando necessário, frente e verso em papel recicável, o qual foi produzido com todos os rascunhos e textos "inutilizados" durante o meu processo de doutoramento. Nesse mesmo contexto, os Apêndices foram disponibilizados eletronicamente.