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

Análise Linguística da Operação de Especificação na Sumarização Humana Multidocumento

Carla Chuman
Ariani Di Felippo

NILC-TR-14-04

Setembro, 2014

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

Resumo

Neste relatório, descreve-se a investigação da especificação no CSTNews, *corpus* multidocumento jornalístico do português. A especificação é das várias operações de fusão *cross-document* identificadas no âmbito da sumarização humana. Com base nela, os humanos expressam no sumário conteúdo mais específico que o ocorrido nos textos-fonte. Essa investigação englobou a (i) descrição das operações de "recorta e cola" (p.ex.: combinação sentencial, expansão sintagmática, substituição sentencial, substituição lexical, etc.) envolvidas no processo de especialização e a (ii) proposição de estratégias para realizar a especialização de forma automática. Com isso, gera-se conhecimento para subsidiar uma futura produção automática de *abtracts* (isto é, sumários elaborados pela reescrita dos textos-fonte). A pesquisa ora descrita foi realizada em uma iniciação científica que compreendeu o período de 01/09/2013 a 31/08/2014.

Este trabalho contou com o apoio financeiro
da FAPESP (2013/12511-0).

1. Introdução

Produzir uma versão condensada (coesa e coerente) de uma coleção de textos de diferentes fontes (de divulgação) que abordam um mesmo assunto é denominada "sumarização multidocumento". Os sumários multidocumento do tipo "informativo e genérico" são os mais tradicionais. Eles são "informativos" porque veiculam a informação central da coleção a ponto de a sua leitura substituir a leitura dos textos-fonte e são "genéricos" porque o seu público alvo não é específico, mas genérico (MANI, 2001; KUMAR, SALIM, 2012).

A produção manual de um sumário, como os multidocumento, é comumente realizada em 3 estágios consecutivos: (i) análise das características dos textos-fonte, (ii) seleção do conteúdo relevante e, por fim, (iii) produção do sumário a partir da materialização linguisticamente do conteúdo selecionado na etapa anterior (CREMMINS, 1996; ENDRES-NIGGEMEYER, 1998).

Para realizar cada um dos estágios, estratégias de interpretação e/ou geração de texto são empregadas pelos humanos. Para condensar o conteúdo selecionado dos textos-fonte, por exemplo, algumas operações de fusão de informação (do inglês, *cross-document fusion*) podem ser empregadas, como eliminação, união, intersecção, generalização/especificação e inferência (MANI, 2001). Para produzir o sumário propriamente dito, o conteúdo condensado pelas operações de fusão pode ser linguisticamente expresso com base em operações de *cut-and-paste* ou "recorta e cola" (de palavras, expressões, etc.) dos próprios textos-fonte, como (i) redução sentencial, (ii) combinação sentencial, (iii) transformação sintática, (iv) paráfrase lexical e (v) reordenação (JING, MACKEOWN, 1999; 2000).

O conhecimento sobre os 3 estágios gerais de sumarização humana e sobre a etapa específica de seleção de conteúdo tem guiado o desenvolvido de sumarizadores automáticos[1] capazes de gerar sumários multidocumento extrativos (isto é, produzidos pela justaposição de sentenças extraídas dos textos-fonte sem qualquer alteração) extratos (p.ex.: CASTRO JORGE, 2010; RIBALDO et al., 2012).

Apesar de a boa coesão e coerência dos sumários multidocumento extrativos, os sumarizadores automáticos estão longe de produzirem *abstracts*, ou seja, sumários elaborados pela reescrita dos textos-fonte. Um dos motivos dessa limitação é a falta de conhecimento sobre as operações de fusão de informação e de "recorta e cola" empregadas na sumarização humana multidocumento (SHM).

Assim, investigou-se a especificação no CSTNews, *corpus* multidocumento jornalístico do português (CARDOSO et al., 2011). Essa investigação englobou a descrição das operações de "recorta e cola" envolvidas no processo e a proposição de estratégias para realizá-las de forma automática.

Na Seção 2, apresentam-se conceitos básicos sobre a sumarização humana e as operações de sumarização e de *cut-and-paste*. Na Seção 3, apresenta-se o *corpus* CSTNews. Na Seção 4, descrevem-se as várias tarefas que compuseram a descrição das operações de "recorta e cola" envolvidas na especialização. Na Seção 5, apresentam-se as estratégias para a realização automática da especialização aqui propostas. Por fim na Seção 6, tecem-se considerações finais sobre as atividades ora relatadas e apontam-se trabalhos futuros.

---

[1] Os sumarizadores automáticos multidocumento são aplicações computacionais desenvolvidas na Sumarização Automática Multidocumento (SAM), uma das subáreas do Processamento Automático de Línguas Naturais (PLN) (MANI, 2001).

2. Revisão da literatura

Especificamente, os seguintes tópicos foram revisados: (i) a sumarização humana (ou manual), (ii) a SHM, (iii) as operações básicas de fusão de conteúdo *cross-document* e (iv) as operações de "recorta e cola" ou reescrita.

2.1. A sumarização humana

A tarefa genérica de sumarização pode ser concebida como a seleção de conteúdo a partir de um ou mais texto-fonte para produzir uma versão condensada do(s) mesmo(s) em função de um usuário ou tarefa (MANI, 2001).

No geral, sabe-se que essa tarefa é bastante subjetiva, pois, sendo uma atividade intelectual, é influenciada pelo conhecimento prévio, atitude e disposição do escritor. Assim, a seleção da informação para produção de um sumário pode depender de: (i) objetivos do autor do sumário, (ii) objetivos ou interesses de seus possíveis leitores e (iii) importância relativa (e subjetiva) que o próprio sumarizador atribui às informações textuais (LUHN, 1958).

Apesar disso, alguns autores, como Cremmins (1996) e Endres-Niggemeyer (1998), identificaram empiricamente que os humanos realizam a sumarização em 3 estágios, descritos no Quadro 1. Endres-Niggemeyer (1990), contudo, ressalta que as estratégias específicas empregadas em cada estágio podem variar em função do indivíduo produtor do sumário.

Quadro 1: Etapas genéricas da sumarização humana.

Estágio	*Meta*	Resultado
Exploração do documento	Identificar as características básicas do material, ou seja, examinar o título do documento, seu perfil, a disposição e formato das informações, a estrutura global e o gênero do documento, familiarizar-se com o conteúdo, etc.	Elaboração de um "esquema" (do ingles, *scheme*), isto é, conhecimento inicial sobre o tipo de documento e estrutura da informação.
Avaliação da relevância	Identificar as unidades textuais relevantes, representar o texto em nível discursivo e combinar os elementos da etapa de exploração, associando-os aos elementos da etapa em questão.	Especificação do "tema" (do inglês, *theme*), ou seja, representação mental estruturada sobre o conteúdo do texto.
Produção do sumário	Transportar itens textuais relevantes do texto-fonte ou original ao sumário e reorganizá-los em uma nova estrutura utilizando, para tanto, as operações de "recorta e cola" ou reescrita.	Produção de uma versão condensada a partir dos pontos mais importantes de um documento.

Além dos fatores mencionados por Luhn (1958), a realização de cada um dos 3 estágios é fortemente influenciada por: (i) o tamanho do sumário, (ii) a audiência, (iii) a função do sumário, (iv) tipo do sumário e (v) o número de textos-fonte e outros (MANI, 2001).

O tamanho desejado do sumário recebe o nome de "taxa de compressão" e dita o quanto de conteúdo deve ser selecionado dos textos-fonte e também qual informação selecionar. Especificamente, um sumário com taxa de compressão de 70% apresenta tamanho equivalente a 30% do tamanho do texto original (em geral,

medido em número de palavras). No cenário multidocumento, a referência para o cálculo da taxa de compressão é o tamanho do maior texto-fonte da coleção.

O tipo de audiência também influencia a produção dos sumários, pois, para um público genérico, produz-se um sumário que veicula a informação principal dos textos-fonte, e, para uma audiência específica, produz-se um sumário que veicula a informação particular de interesse desse público.

Quanto à função do sumário, ressalta-se que este pode ser (i) informativo, isto é, conter as informações principais do texto-fonte ao ponto de dispensar a leitura do original (p.ex.: *abstracts* de artigos científicos), (ii) indicativo, ou seja, veicular os tópicos do texto-fonte, normalmente exibidos em lista (p.ex.: índices de livros) ou (iii) críticos, isto é, apresentar, além da informação principal do texto-fonte, opinião crítica a respeito dele (p.ex.: resenhas de livros e filmes).

Quanto ao tipo ou forma, os sumários podem ser extrativos, ou seja, formados por trechos (p.ex.: sintagmas e sentenças) extraídos integralmente dos textos-fonte, ou abstrativos, isto é, compostos por material linguístico que foi reescrita em função do texto-fonte. Especificamente, os sumários gerados de forma extrativa são denominados simplesmente de "extratos" e os construídos de forma abstrativa, de *abstracts*.

Quanto à quantidade de textos-fonte, a sumarização pode ser monodocumento ou multidocumento. Na sumarização monodocumento, os humanos produzem um sumário a partir de apenas um texto-fonte. Nesse cenário, os desafios englobam principalmente a identificação da informação relevante do documento-fonte em função do tipo de sumário e da audiência e a produção de uma versão condensada que seja coerente e coesa (MANI, 2001). Sobre a SHM, foco deste trabalho, informações mais detalhadas são apresentadas na próxima subseção.

2.2. A sumarização humana multidocumento

Na SHM, o humano produz um resumo a partir de uma coleção de textos (cada um deles proveniente de uma fonte distinta) que abordam um mesmo assunto.

Apesar de pouco intuitiva, a sumarização multidocumento é relativamente comum em alguns domínios (MANI, 2001). No âmbito do jornalismo, pode-se ditar o "*clipping*" como um exemplo de sumarização multidocumento. O *clipping* é processo bastante antigo na área da comunicação e se define pela pesquisa e seleção contínua de notícias relacionadas a certos assuntos, com o objetivo de atender a um público direcionado (TEIXEIRA, 2001). Segundo Teixeira (2001), o *clipping* de mídia impressa enquanto produto do processo de pesquisa e seleção pode ser materializado nas formas denominadas: (i) clássica, ou seja, um conjunto de recortes de notícias, reportagens, artigos, etc., (ii) sinopse, isto é, um texto que contempla as principais notícias de interesse do cliente, ou (iii) análise, isto é, um texto que contém uma interpretação crítica das informações coletadas. Atualmente, tem-se em foco o *clipping* eletrônico (*e-clipping* ou *web clipping*), que consiste em selecionar, coletar e organizar as informações veiculadas por diversas fontes da *web* a respeito de certo assunto (pessoa, evento, instituição etc).

Os resultados da seleção podem ser organizados e veiculados não só nas formas clássicas, sinopse ou análise, mas também nos formatos de: (i) lista de *hyperlinks*, em que cada *link* leva a um *site* ou documento específico, e (iii) lista de excertos de documentos. Hoje, há várias empresas especificadas em *clipping* eletrônico, como a Associação Brasileira das Empresas de Monitoramento de

Informação (ABEMO)[2], o Grupo Info4[3], o Armazém Digital[4], entre outras. E, além das empresas, inúmeras instituições e associações também oferecem *clippings* eletrônicos a seus associados e/ou consulentes, como a Universidade Federal de São Carlos (UFSCar)[5], entre outras.

No mercado editorial, os sumários multidocumento constituem, por exemplo, as introduções de coletâneas de artigos e de livros, nas quais as informações principais de cada artigo ou capítulo são fornecidas aos leitores.

Além dos fatores gerais que afetam a sumarização, a produção de um sumário a partir de uma coleção de textos que tratam de um mesmo assunto é afetada por outros fenômenos.

Especificamente na fase de "exploração do(s) documento(s)" (cf. Quadro 2), o humano realiza a comparação dos textos-fonte com o objetivo de identificar conteúdo repetido, complementar e contraditório na coleção. Nesse processo, o produtor do sumário precisa lidar ainda com outros fenômenos típicos da multiplicidade de textos-fonte, como estilos de escrita variados, ordenação temporal dos eventos/fatos e perspectivas e focos distintos, os quais ocorrem porque os documentos têm origem diversificada e são escritos em diferentes momentos.

Após a "exploração do(s) documento(s)", o humano precisa identificar o conteúdo relevante para compor o sumário. Esse processo de seleção é bastante complexo porque é influenciado pelo domínio que o produtor tem do assunto tratado nos textos. Apesar disso, alguns trabalhos comprovaram que os humanos concordam sobre a informação principal, delimitando-a com base na redundância do conteúdo na coleção (MANI, 2001; NENKOVA, 2006; CAMARGO, 2013).

Nenkova (2006), por exemplo, constatou essa característica ao verificar que as sentenças que compõem esse tipo de sumário apresentam as palavras (de conteúdo) mais frequentes da coleção. Utilizando um *corpus* em inglês de 30 coleções, cada uma composta por 10 textos jornalísticos compilados do jornal *The New York Times*, 10 sujeitos produziram um sumário multidocumento com aproximadamente 100 palavras para cada coleção. Desse experimento, a autora comprovou que em média 94,66% das palavras mais frequentes de uma coleção estão presentes nos respectivos sumários.

Camargo (2013), por sua vez, constatou a importância da redundância como critério de seleção de conteúdo ao verificar que o conteúdo das sentenças dos textos-fonte (de dada coleção) que possuem maior número de relações semânticas estabelecidas entre si entre comumente é selecionado para compor o sumário.

Para condensar o conteúdo selecionado, os humanos fazem uso de algumas operações básicas de fusão de informação, como eliminação, união, intersecção, generalização/especificação e inferência (MANI, 2001).

Na produção do sumário, o conteúdo condensado pelas operações de fusão pode ser linguisticamente expresso com base em operações de "recorta e cola" ou reescrita (de palavras, expressões, etc.) dos próprios textos-fonte, como (i) redução sentencial, (ii) combinação sentencial, (iii) transformação sintática, (iv) paráfrase lexical e (v) reordenação (JING, MACKEOWN, 1999; 2000).

As operações de fusão de informação e de reescrita dos textos-fonte são descritas nas subseções (c) e (d), respectivamente.

---

[2] http://www.abemo.org/
[3] http://www.info4.com.br/info4/novosite/site2/
[4] http://www.adigital.com.br/
[5] http://www.ccs.ufscar.br/clipping

2.3. As operações de fusão de conteúdo *cross-document*

A eliminação, também denominada deleção, é a operação de condensação em que certas informações contidas nos textos-fonte são removidas ou excluídas.

A união é uma operação por meio da qual as informações distintas, provenientes de textos distintos, são preservadas no sumário, eliminando-se a redundância.

A intersecção consiste em combinar informações que se repetem nos textos-fonte para produção do sumário.

A generalização, por sua vez, é o processo no qual determinado conteúdo dos textos-fonte é expresso no sumário por um elemento textual mais genérico ou abstrato. Caso o conteúdo do sumário for mais específico que dos textos-fonte, tem-se a especificação, foco deste trabalho.

Para exemplificar tais operações, consideram-se as sentenças-fonte apresentadas em (1). Especificamente, há, em (1), 3 sentenças de documentos-fonte (D) distintos que veiculam mesma notícia. No caso, essas sentenças são redundantes entre si, pois veiculam conteúdo similar.

1) O Airbus A320, voo JJ3054, partiu de Porto Alegre, às 17h16 da terça-feira e chegou São Paulo às 18h45. (D1)

   A aeronave da TAM Airbus A320, voo JJ3054, partiu de Porto Alegre, às 17h16 com destino a Congonhas. (D2)

   Um Airbus A320 com capacidade para 170 passageiros partiu de Porto Alegre (RS) às 17h16 com destino a Congonha. (D3)

Com base na operação de fusão do tipo **união**, gera-se, por exemplo, o sumário em (2), constituído por apenas uma sentença. Nela, as informações distintas provenientes de cada uma das sentenças-fonte são preservadas, de tal forma a não apresentar redundância.

2) A aeronave da TAM Airbus A320, voo JJ3054, com capacidade para 170 passageiros, partiu de Porto Alegre (RS), às 17h16 da terça-feira com destino a Congonhas e chegou a São Paulo às 18h45.

Ao se considerar os 3 textos-fonte que compõem o exemplo em (1), é possível observar as informações de (2) que são comuns a mais de um dos textos-fonte e as informações particulares de cada um dos textos-fonte. Os itens de (a) e (f), a seguir, explicitam a origem dos diferentes trechos que compõem a sentenças-resumo em (2).

a) D2 → A aeronave da TAM
b) D1, D2, D3 → Airbus A320 / partiu de Porto Alegre / às 17h16
c) D1, D2 → voo JJ3045
d) D3 → com capacidade para 170 passageiros / (RS)
e) D1 → da terça-feira / chegou a São Paulo às 18h45
f) D2, D3 → com destino a Congonhas

Sobre os item (a-f), observa-se que o trecho em (a) (“A aeronave da TAM”) é proveniente exclusivamente do texto-fonte D2, o mesmo acontece com os trechos em (d) e (e), que são originários dos documentos D3 e D1, respectivamente. Os trechos em (a), (d) e (e) foram unificados para compor o sumário em (2). E as informações

comuns a mais de um texto, como as em (b), (c) e (f) foram trazidas para o sumário, eliminando-se a redundância.

Com base na operação de fusão do tipo **intersecção**, por sua vez, é possível gerar o a sentença-sumário em (3). Nela, somente as informações que mais se repetem nos textos-fonte (ou seja, as informações redundantes) estão combinadas. Especificamente, apenas os trechos dos itens (b), (c) e (f) foram selecionados das sentenças-fonte em (1) e combinados para a produção da sentença-resumo em (3).

3) O Airbus A320, voo JJ3054, partiu de Porto Alegre, às 17h16 com destino a Congonhas.

Ademais, observa-se que a produção de (3) englobou não só a intersecção das informações mais redundantes, mas também a deleção ou eliminação das informações específicas de cada texto veiculadas pelos trechos abaixo:

(i) "A aeronave da TAM" (ou seja, a companhia de voo), de D2;
(ii) "com capacidade para 170 passageiros" (ou seja, capacidade do avião), de D3,
(iii) "(RS)" (ou seja, o estado de partida do voo), de D3;
(iv) "da terça-feira" (ou seja, dia da semana em que o voo partiu), de D1;
(v) "chegou a São Paulo às 18h45" (ou seja, local e horário de chegada do voo), de D1.

A generalização é uma operação de condensação bastante utilizada para reduzir o conteúdo dos textos-fonte e, por conseguinte, produzir um sumário correspondente. Com base na generalização, o conteúdo dos textos-fonte é expresso de forma mais abstrata ou genérica. Para tanto, os sumarizadores humanos também precisam realizar um processo de inferência. Aplicando-se a generalização às sentenças-fonte do exemplo (1), pode-se produzir, por exemplo, o sumário em (4).

(4) O avião partiu de Porto Alegre com destino a São Paulo.

Nele, notam-se 2 casos de **generalização**.

Um deles diz respeito ao fato de que o objeto "A aeronave da TAM Airbus A320" (D1 e D2) foi generalizado para "avião". Essa generalização envolveu a deleção das informações sobre seu modelo ("Airbus A320"), número ("voo JJ3045"), capacidade de carga ("com capacidade para 170 passageiros") e companhia ("TAM"). A utilização no sumário do item lexical "avião" também pode ser uma espécie de generalização, pois "avião" é uma palavra menos técnica que "aeronave".

O outro caso de generalização refere-se ao evento principal ("a partida do voo"), que implicou na deleção do (i) horário e dia da partida do voo e (ii) do aeroporto de destino.

Por fim, a **especificação**, foco deste trabalho, é outra operação de fusão *cross-document*. Com base nela, expressa-se no sumário conteúdo mais específico que o ocorrido nos textos-fonte, resultante também de um processo de inferência. A partir das mesmas sentenças-fonte de (1), pode-se gerar por meio de inferências, por exemplo, uma sentença-sumário com informações mais específicas que as veiculadas nas sentenças-fonte, como a ilustrada em (5).

5) O Airbus modelo A320, voo JJ3054, partiu de o aeroporto Salgado Filho, em Porto Alegre, às 17h16 com destino a Congonhas.

Nota-se em (5) que o sumário apresenta duas informações mais específicas que as das sentenças que lhe deram origem (1). Com base principalmente em conhecimento de mundo, o produtor do sumário explicitou que "A320" trata-se de um "modelo" de Airbus e que a origem do voo, por ser Porto Alegre, trata-se do aeroporto Salgado Filho.

A seguir, descrevem-se as principais operações de reescrita por meio das quais as operações de fusão de conteúdo se materializam nos sumários.

2.4. As operações de *cut-and-past* ou "recorta e cola" dos textos-fonte

Do ponto de vista da materialização linguística, o conteúdo condensado pelas operações de fusão *cross-document* pode ser expresso no sumário com base em estratégias de recorta e cola ou reescrita de diferentes segmentos (p.ex.: palavras, expressões, etc.) dos próprios textos-fonte, como evidenciam Jing e Mackeon (1999, 2000). As principais operações da literatura são:

(i) **redução sentencial**: esse é o processo pelo qual a sentença de um sumário é escrita a partir da remoção de palavras, expressões e sintagmas de sentenças dos textos-fonte;
(ii) **combinação sentencial**: por meio desse processo, a sentença de um sumário multidocumento é formulada pela combinação de elementos dos textos-fonte;
(iii) **transformação sintática**: esse é o processo pelo qual a sentença de um sumário é formulada com base na transformação sintática das sentenças dos textos-fonte (p.ex.: mudança de posição do sujeito, mudança de posição das palavras ou sintagmas); esse processo engloba outro, o de reordenação;
(iv) **paráfrase lexical**: processo que se caracteriza pala alteração de vocábulos dos textos-fonte por sinônimos ou expressões equivalentes.

Na sentença-resumo em (2), por exemplo, o conteúdo condensado pela união foi expresso a partir da reescrita do material linguístico superficial dos textos-fonte. Essa reescrita se deu pelo emprego da **combinação sentencial**, já que a sentença-sumário combina trechos advindos das diversas fontes.

Já em (3), observa-se que a condensação ocorreu pela **redução sentencial**, já que os elementos "A aeronave da TAM" (texto 2), "com capacidade para 170 passageiros" e "(RS)" (texto 3) e "às 17h16" (textos 1, 2 e 3) e "da terça-feira" (texto 1) foram desconsiderados no sumário. Aliás, os elementos deletados são de diferentes granularidades, por exemplo: (RS) = sigla e "da terça-feira" = sintagma preposicional.

Em (4), o conteúdo condensado pela operação de generalização foi expresso principalmente pelas estratégias de **redução sentencial**, já ilustrada, e **paráfrase lexical**, que se caracterizou pela escolha da unidade lexical "avião" (de língua geral) em detrimento de unidades mais técnicas, como "aeronave" ou "Airbus".

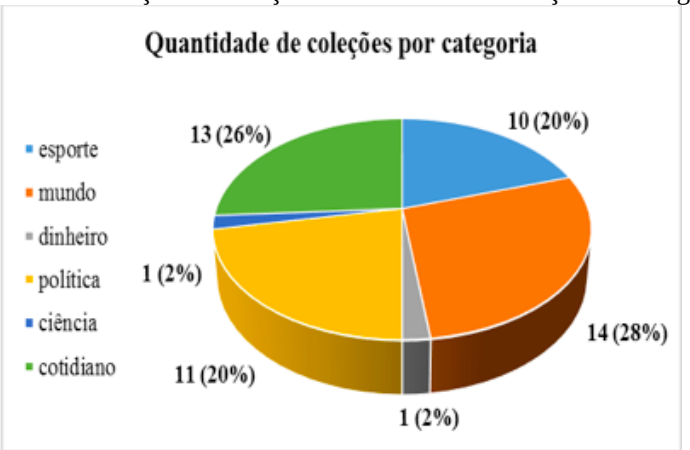
A partir da revisão da literatura apresentada, adquiriu-se o arcabouço teórico para a investigação ora proposta. Em especial, a revisão permitiu compreender mais a operação de especificação do conteúdo original para a produção de sumários multidocumento e, sobretudo, elencar as principais operações de "recorta e cola" utilizadas para materializar a especificação.

As operação de “recorta e cola” ou reescrita identificadas na literatura foram o ponto de partida para a caracterização dos casos de especificação do *corpus* multidocumento de referência do português, o CSNews (CARDOSO et al., 2011), que é descrito na sequência.

3. Seleção e Recorte do *corpus*

Para investigar a operação de especificação no cenário multidocumento, selecionou-se o CSTNews (CARDOSO et al., 2011), *corpus* multidocumento de referência em português do Brasil. O CSTNews é composto por 50 coleções de textos jornalísticos. Cada uma dessas coleções engloba textos que versam sobre um mesmo assunto. Os textos-fonte são do gênero “notícias jornalísticas”, que, segundo alguns autores (p.ex.: BARBOSA (2001), DOLZ; SCHNEWLY (2004), LAGE (2002, 2004)), caracterizam-se por: (i) documentar as experiências humanas vividas (domínio social) e (ii) representar pelo discurso as experiências vividas, situadas no tempo (capacidade da linguagem). Especificamente, cada coleção do CSTNews contém basicamente: (i) 2 ou 3 textos sobre um mesmo assunto, compilados de diferentes fontes jornalísticas e com tamanhos similares; (ii) sumários humanos monodocumento e multidocumento; (iii) sumários automáticos multidocumento, gerados pelos sumarizadores automáticos para o português de melhor desempenho; (iv) anotações linguísticas monodocumento em diferentes níveis (morfossintática, sintática, semântica e semântico-discursiva) e multidocumento em nível semântico-discursiva. Os textos-fonte *corpus* foram manualmente coletados dos jornais *online* *Folha de São Paulo, Estadão, Jornal do Brasil, O Globo* e *Gazeta do Povo*. As coleções possuem em média 42 sentenças (de 10 a 89) e os sumários, 7 sentenças em média (de 3 a 14). Ademais, as coleções estão rotuladas pelas “seções” dos jornais dos quais os textos foram compilados. Assim, o *corpus* é composto por coleções das categorias: “esporte”, “mundo”, “dinheiro”, “política”, “ciência” e “cotidiano” (Figura 1).

Figura 1: Distribuição das coleções do CSTNews em função das categorias.



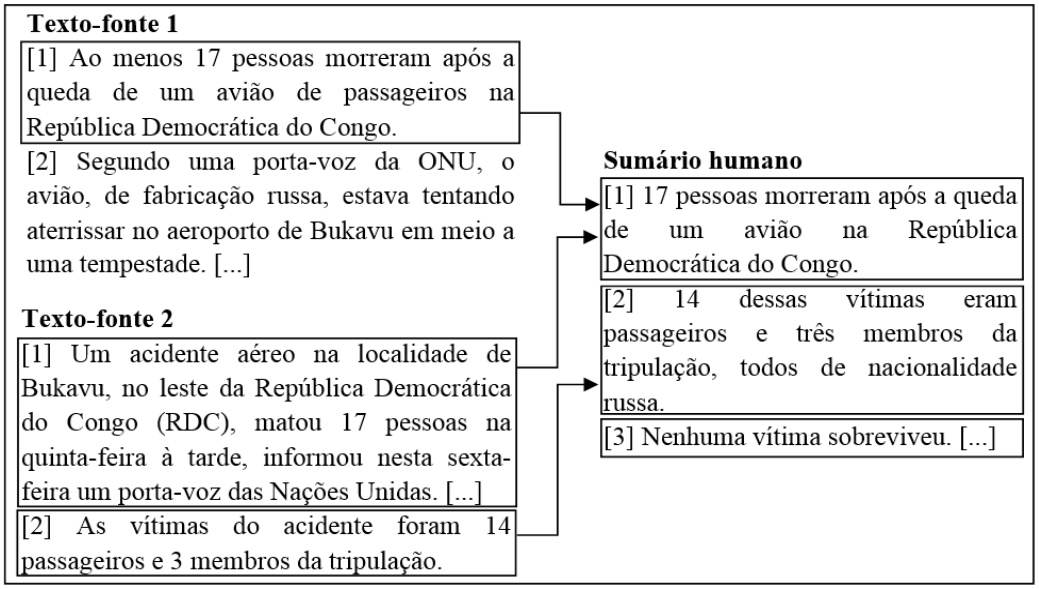
Quanto aos sumários multidocumento, em especial, ressalta-se que eles são informativos e genéricos e foram construídos manualmente de forma abstrativa, ou seja, com reescrita dos textos-fonte. Além disso, a produção dos mesmos foi guiada por uma taxa de compressão de 70%. Consequentemente, os sumários contêm, no máximo, 30% do número de palavras do maior texto-fonte da coleção.

Com o objetivo de investigar o processo de SHM, as sentenças do sumário multidocumento de cada coleção do CSTNews foram manualmente alinhadas a todas as sentenças dos documentos-fonte de origem (AGOSTINI et al., 2012; CAMARGO, 2013).

Na Figura 2, retirada de Camargo et al. (2011), ilustram-se três alinhamentos entre o sumário da coleção C1 do CSTNews, que pertence à categoria "mundo", e seus respectivos documentos ou textos-fonte. Observa-se, no exemplo, que a sentença [1] do sumário foi alinhada à sentença [1] do Texto-fonte 1 e às sentenças [1] e [2] do Texto-fonte 2. No caso da coleção C1, nenhuma sentença do sumário foi alinhada à sentença [2] do Texto-fonte 1.

O alinhamento entre o sumário e seus respectivos textos-fonte em cada coleção foi feito com base em duas regras gerais. A primeira delas diz respeito ao nível dos segmentos a serem alinhados e a segunda refere-se ao critério para a identificação das correspondências.

Figura 2: Exemplo de alinhamento sumário/textos-fonte do CSTNews.



Quanto ao nível dos segmentos textuais, Agostini et al. (2012) optaram pelo alinhamento em nível sentencial, pois as sentenças são unidades informacionais bem delimitadas. Sobre o critério de alinhamento, ressalta-se que as correspondências entre as sentenças dos sumários (SS) e as sentenças dos documento-fonte (SD) foram identificadas com base na sobreposição de conteúdo, total ou parcial. Tendo em vista que os sumários do CSTNews são *abstracts*, o alinhamento nem sempre foi simples, pois, pela reescrita do conteúdo dos textos-fonte no sumário, a identificação da origem da informação não é trivial.

Ao se optar por um alinhamento manual baseado na sobreposição de conteúdo ou informação, o processo de indexação não se baseia apenas na sobreposição de formas (ou seja, de unidades lexicais), como acontece quando esse alinhamento é feito de forma automática. Consequentemente, sentenças que continham conteúdo em comum, seja com alta (6) ou baixa sobreposição lexical (7), foram alinhadas. Nos exemplos (6) e (7), o sublinhado indica as palavras em comum entre SS e SD.

Por exemplo, calculando-se a sobreposição de conteúdo entre as sentenças de (6) por meio da medida estatística *word overlap*[6], que se baseia exclusivamente da ocorrência de palavras iguais, obtém-se o valor 0.8, que indica alta sobreposição. Ao calcular a *word overlap* entre as sentenças de (7), obtém o valor 0.38, que indica baixa sobreposição lexical. Mesmo com essa diferença, as sentenças foram alinhadas em decorrência da metodologia de alinhamento ser sido manual e da sobreposição de conteúdo como critério básico para conexão das sentenças dos sumários às suas sentenças de origem nos textos-fonte.

6)

SS: O próximo objetivo da seleção é a medalha de ouro nos Jogos Pan-Americanos do Rio.

SD: Seu próximo objetivo é os Jogos Pan-Americanos do Rio.

7)

SS: A pista principal do Aeroporto Internacional de São Paulo (Cumbica), em Guarulhos, será totalmente reformada a partir de março de 2008, segundo informações do Ministério da Defesa.

SD: O Ministério da Defesa anunciou nesta segunda-feira (6) que em março do ano que vem uma das pistas do Aeroporto de Guarulhos será fechada para reformas de seu trecho central.

Os alinhamentos do CSTNews também foram manualmente tipificados. Dado um par de sentenças alinhadas, a conexão entre elas foi caracterizada em função da sobreposição de material linguístico, da operação de condensação e da sua natureza onomástica.

Quanto à sobreposição de forma, os alinhamentos foram anotados com as etiquetas: (i) idêntico, (ii) parcial ou (iii) diferente.

Sobre as operações de fusão, os alinhamentos receberam as etiquetas: (i) especificação, (ii) generalização, (iii) inferência, ou (iv) neutro. O rótulo "neutro" foi utilizado nos casos em que o conteúdo de uma SD foi transposto à SS por alguma operação que não as codificadas pelas outras etiquetas (i-iii). Além disso, os alinhamentos foram caracterizados em função de sua natureza onomástica (isto é, relativa a nomes próprios), pois as informações veiculadas por nomes próprios são alvo frequentes das operações de condensação *cross-document*. Assim, os alinhamentos foram anotados com as etiquetas: (i) toponímia, quando a operação se deu sobre nomes de lugares, ou (ii) antroponímia, quando a operação foi feita sobre nomes de pessoas.

Em (8), por exemplo, tem-se um alinhamento tipificado com as etiquetas: **parcial** e **especificação**. Nesse exemplo, a SS foi alinhada à SD por expressar parcialmente o conteúdo da SD. O conteúdo parcial diz respeito à informação de "quando ocorreu o primeiro gol do Brasil". Pelo fato de esse alinhamento ter sido anotado com a etiqueta "especificação", a SS expressa essa informação de forma

[6] Para calcular a *word overlap* entre um par de sentenças (S1 e S2), aplica-se a fórmula: CommonWords / (Words (S1) + Words (S2)). Assim, divide-se o número total de palavras idênticas entre as sentenças (CommonWords) pela soma do número total de palavras de cada sentença (Words(S1) + Words(S2)), excluindo-se as *stopwords* (palavras de classe fechada), números e símbolos. O resultado obtido será entre 0 e 1, sendo que, quanto mais próximo de 1 for a Wol, mais redundante será o par entre si, e, quanto mais próximo de 0, menos redundante (JURAFSKY, MARTIN, 2001).

mais específica. No caso, a informação "início da partida", de SD, foi especificada para "primeiros 4 minutos do jogo". Os trechos específicos envolvidos no caso de especificação estão sublinhados em (8).

8) **SS**: É verdade que o Brasil deu sorte de conseguir um gol logo no início da partida.
SD: O Brasil conseguiu um gol logo nos primeiros 4 minutos do jogo, fazendo os argentinos apertarem o ataque no jogo, restando ao Brasil os contragolpes, chegando ao segundo gol, que foi um gol contra.

Quanto à sobreposição de forma, ressalta-se que, dos 1007 alinhamentos manuais do CSTNews: (i) 949 foram rotulados como neutros, o que representa 94,2% dos casos; (ii) 82 alinhamentos foram caracterização como generalizações, o que totaliza 8.1% dos alinhamentos, (iii) **47** alinhamentos receberam a etiqueta "**especificação**", que representa 4.7% do total de alinhamentos, e (iv) 33 alinhamentos englobam SSs que resultaram especificamente de inferências a partir das SDs (3.2%) (AGOSTINI *et al.*, 2012). Especificamente quanto à especificação, 36 sentenças distintas de diferentes sumários foram conectadas a suas sentenças de origem, gerando o total de 47 alinhamentos nos quais o processo de especificação foi observado e anotado.

Os 47 casos do CSTNews foram manualmente revisados, o que levou à exclusão de 8 alinhamentos (envolvendo 5 SSs) por não englobarem efetivamente processos de especificação.

Consequentemente, das 36 SSs alinhadas, restaram 31 e, dos 47 alinhamentos, 39 foram efetivamente alvo da investigação sobre o processo de especificação.

Ressalta-se que, entre os 39 alinhamentos restantes, alguns deles se caracterizam pela conexão de uma SS a apenas uma SD (ou seja, alinhamento do tipo 1-1) e outras se caracterizam pela conexão de uma SS a mais de uma SD (ou seja, alinhamento do tipo 1-*n*). A distribuição dos diferentes tipos de alinhamento com especificação do CSTNews está descrita na Tabela 1. No caso, observa-se que os alinhamentos concentram-se nos tipos 1SS→1SD e 1SS→2SD apenas.

Tabela 1: Quantificação numérica dos tipos de alinhamento com especificação.

	Tipos de alinhamento	
	1-1	1-2
Número de alinhamentos	23	8

Em outras palavras, pode-se dizer que cada uma das 31 sentenças dos sumários envolvidas nos alinhamentos do tipo simples (ou seja, 1-1) foi elaborada a partir do conteúdo e do material linguístico de apenas uma sentença, ao passo que cada SS nos alinhamentos complexos (no caso, 1-2) foi construída a partir de conteúdo e material de linguístico de duas SD.

Na sequência, descreve-se a tarefa de identificação das operações de "recorta e cola" ou reescrita envolvidas nos 39 casos de especificação selecionados do CSTNews.

4. Identificação das operações de "recorta e cola"

A identificação das operações de "recorta e cola" foi feita em 3 passos. O primeiro passo consistiu na organização e registro dos dados para análise. O segundo deles consistiu na identificação, delimitação e indexação dos casos de especificação nos 39 alinhamentos. E, por fim, o terceiro passo englobou a identificação, anotação e descrição das operações de "recorta e cola" ou reescrita. Para tanto, tomou-se como ponto de partida o conjunto inicial de operações da literatura.

4.1. Organização e registro dos dados para análise

Todo o processo de identificação das operações de "recorta e cola" referentes à especificação no CSTNews foi feito por meio do *Microsoft Excel*. Especificamente, criou-se uma tabela no formato *Excel* como a ilustrada no Quadro 2.

Por meio do Quadro 4, observa-se que, na coluna A da tabela, estão registados os casos em que uma SS foi alinhada a uma ou mais SDs, totalizando 31. No Quadro 4, tem-se os casos 1 e 7.

Na coluna B, tem-se a referência do *cluster* ou coleção a que pertencem as sentenças alinhadas. O caso 1 foi observado no *cluster* C2 e o caso 7, no *cluster* C22.

Na coluna C, descreve-se a SS e a informação, entre parêntese, da localização de ocorrência da mesma no sumário. Por exemplo, a SS do caso 1 é "*Quando se compara com uma pesquisa sem a lista oficial dos candidatos, Lula sobe de 27% para 31%, Geraldo de 4% a 14% e Heloisa de 1% a 6%.*", que ocorreu na posição 7.

Na coluna D, tem-se as sentenças dos documentos-fonte às quais a SS foi alinhada. No Quadro 4, a SS do caso 1 foi alinhada a uma única SD; trata-se da S6 do D1 do *cluster* C2 ("*A pesquisa de hoje apresentou uma variação na lista espontânea (quando os entrevistados dizem em quem pretendem votar sem um cartão com nomes).").*

A coluna E foi destinada ao registro de interpretações ou comentários sobre os casos de especificação e as operações de reescrita.

A	B	C	D	E
Caso	**Cluster**	**SS**	**SD**	**Comentário**
1	C2	Quando se compara com uma pesquisa sem a lista oficial dos candidatos, Lula sobe de 27% para 31%, Geraldo de 4% a 14% e Heloisa de 1% a 6%. (S7)	A pesquisa de hoje apresentou uma variação na lista espontânea (quando os entrevistados dizem em quem pretendem votar sem um cartão com nomes). (S6_D1)	
7	C22	Depois de uma segunda-feira caótica, em que 38,1% dos vôos atrasaram em todo o país, o aeroporto de Congonhas esteve fechado nesta manhã de terça-feira, devido ao nevoeiro que cobria a região sul de São Paulo (S1)	Depois de uma segunda-feira caótica, quando 38,1% dos vôos atrasaram em todo o País, a terça-feira deve ser mais um dia de longas filas nos aeroportos por conta dos problemas em Congonhas. (S6_D2)	
			O aeroporto esteve fechado devido ao mau tempo (S5_D4)	

Quadro 2: Organização dos dados para análise.

Na sequência, descrevem-se as tarefas de (i) identificação, delimitação e indexação dos casos de especificação e (ii) identificação, anotação e descrição das operações de "recorta e cola". Ao final da descrição de cada uma dessas tarefas, o Quadro 3 será enriquecido com os dados provenientes de cada uma delas. Ao final desta seção, ter-se-á um quadro com o resultado final das análises.

4.2. Identificação, delimitação e indexação dos casos de especificação

Os 39 alinhamentos do CSTNews foram manualmente analisados de forma individual com o objetivo de identificar efetivamente os casos de especificação, pois uma SS associada a uma única SD pode apresentar mais de um caso de especificação, assim como uma SS alinhada a várias SDs pode conter apenas uma especificação.

Essa identificação englobou a delimitação e a indexação dos trechos especificados nas SSs e dos trechos de origem nas SDs. Essas tarefas se mostraram pertinentes porque os alinhamentos do CSTNews foram feitos sem a delimitação explícita dos trechos das sentenças envolvidos nos processos de fusão *cross-document* como a especificação.

No Quadro 3, ilustram-se a delimitação dos trechos sentenciais por colchetes "[ ]", que marcam o início e o fim de cada um deles, e a conexão do(s) trecho(s) de uma SS ao(s) trecho(s) da(s) SD(s) por índices numéricos. Por exemplo, a S7 do sumário (C2), alinhada à S6 do D1, tem 1 caso de especificação delimitado e indexado.

A S1 do C22, por sua vez, possui 2 casos de especificação. Especificamente, registrou-se de forma explícita que o alinhamento da SS à SD foi rotulado como "especificação" no CSTNews porque a SS possui 2 casos de especificação. Em um desses casos, o trecho [manhã de terça-feira] da SS foi interpretado como especificação de [terça-feira] da S6 do D1, por isso, ambos foram anotados com o índice ("1").

A	B	C	D	E
Caso	**Cluster**	**SS**	**SD**	**Comentário**
1	C2	Quando se compara com uma pesquisa sem a lista oficial dos candidatos, **[Lula sobe de 27% para 31%, Geraldo de 4% a 14% e Heloisa de 1% a 6%.]1** (S7)	A pesquisa de hoje apresentou **[uma variação na lista espontânea]1** (quando os entrevistados dizem em quem pretendem votar sem um cartão com nomes). (S6_D1)	
7	C22	Depois de uma segunda-feira caótica, em que 38,1% dos vôos atrasaram em todo o país, o aeroporto de Congonhas esteve fechado nesta **[manhã de terça-feira]1**, devido ao **[nevoeiro que cobria a região sul de São Paulo.]2** (S1)	Depois de uma segunda-feira caótica, quando 38,1% dos vôos atrasaram em todo o País, **[a terça-feira]1** deve ser mais um dia de longas filas nos aeroportos por conta dos **[problemas em Congonhas.]2** (S6_D2)	
			O aeroporto esteve fechado devido ao **[mau tempo]2** (S5_D4)	

Quadro 3: Delimitação e indexação dos trechos especificados e de origem.

Quanto à natureza dos trechos delimitados, ressalta-se que alguns deles consistem em unidades lexicais ou siglas (p.ex.: [ACM] em (9)) ou sintagmas (p.ex.: [o parlamentar] em (9)). Na maioria dos casos, no entanto, os trechos delimitados consistiram de orações ou sentenças inteiras, como ilustrado em (10).

9) **SS:** Em maio, [ACM]1 passou mal no Senado e chegou a desmaiar em frente ao seu gabinete. (S13)
SD: No final de maio, **[o parlamentar]1** sentiu-se mal no Senado e chegou a cair em frente ao seu gabinete. (S16_D1)

10) **SS:** Segundo Lula, **[o mundo precisa de uma nova matriz energética, e o etanol pode ser mais que uma alternativa de energia limpa]1** (S3)
SD: Precisamos avaliar o caminho percorrido e estabelecer **[novas metas]1** (S3_D4)

Ao total, identificaram-se 45 casos de especificação. Especificamente, das 31 SS alinhadas por especificação, 23 delas possuí apenas 1 caso, 8 delas possuí 2 casos e 2 delas possui 3 casos de especificação.

4.3. Identificação, anotação e descrição das operações de "recorta e cola"

Para essas tarefas, partiu-se das operações identificadas na literatura (redução sentencial, combinação sentencial, transformação sintática e paráfrase lexical). No entanto, das 4 operações da literatura, apenas a combinação sentencial e a transformação sintática ocorreram nos casos de especificação do CSTNews.

Na análise dos casos, foram identificadas outras 6 operações, denominadas: expansão sintagmática, expansão lexical, substituição lexical, substituição sintagmática, substituição sentencial e substituição numérica.

Para cada uma delas, propôs-se uma etiqueta ou *tag* para anotar os trechos das SSS e das SDs delimitados e indexados. Tais etiquetas foram compostas pelas letras iniciais em maiúsculo de cada palavra que constitui a denominação da operação. Para "combinação sentencial", por exemplo, a etiqueta especificada foi **CS**. Para diferenciar os tipos "substituição sintagmática" e "substituição sentencial", foram propostas etiquetas secundárias em minúsculo, a saber: SS**e**, para "substituição sentencial", e SS**i**, para "substituição sintagmática".

O formato de anotação propriamente dito foi: todas as operações de "recorta e cola" referentes ao caso de especificação delimitado e indexado foram agrupadas ao fim, delimitadas entre si pela barra de divisão (/). Assim, o formato genérico da anotação para *n* operações foi: [trecho]índice_operação1/operação2/…/operação_*n*.

Após a anotação das operações de reescrita por meio das quais o conteúdo das sentenças-fonte foi especificado nas sentenças dos sumários, especificou-se um comentário, também indexado, a respeito da operação anotada. Esse comentário funciona como uma espécie de registro da interpretação da especificação que levou à anotação da operação em questão.

No Quadro 4, ilustra-se a anotação e o registro dos comentários.

A	B	C	D	E
Caso	**Cluster**	**SS**	**SD**	**Comentário**
1	C2	Quando se compara com uma pesquisa sem a lista oficial dos candidatos, [Lula sobe de 27% para 31%, Geraldo de 4% a 14% e Heloisa de 1% a 6%.]**1_CS/SSe** (S7)	A pesquisa de hoje apresentou [uma variação na lista espontânea]**1_CS/SSe** (quando os entrevistados dizem em quem pretendem votar sem um cartão com nomes). (S6_D1)	1_CS/SSe: Combinação sentencial (com outra S), explicitação de nome de pessoa e explicitação de dado numérico (porcentagem).
7	C22	Depois de uma segunda-feira caótica, em que 38,1% dos vôos atrasaram em todo o país, o aeroporto de Congonhas esteve fechado nesta [manhã de terça-feira]**1_CS/ES**, devido ao [nevoeiro que cobria a região sul de São Paulo.]**2_CS/SSe** (S1)	Depois de uma segunda-feira caótica, quando 38,1% dos vôos atrasaram em todo o País, a [terça-feira]**1_CS/ES** deve ser mais um dia de longas filas nos aeroportos por conta dos [problemas em Congonhas.]**2_CS/SSe** (S6_D2)	1_CS/ES: Combinação sentencial (com outra S) e expansão sintagmática pela especificação do período do dia. 2_CS/SSe: Combinação sentencial (com outra S) e substituição de dois sintagmas por uma oração "*nevoeiro que cobria a região sul de São Paulo*".
			O aeroporto este fechado devido ao [mau tempo]**2_CS/SSe**	

Quadro 4: Anotação das operações de reescrita e registro de comentários.

Ao final da anotação, 8 operações distintas foram encontradas no *corpus.* A seguir, no Quadro 6, cada uma delas é definida e exemplificada com ocorrências do *corpus*.

Operação	**Combinação sentencial**	
Etiqueta	CS	
Definição	Operação pela qual uma SS (ou um trecho dela) expressa conteúdo especificado resultante da combinação de material linguístico de duas ou mais SDs; as SDs não necessariamente são as que estão alinhadas à SS.	
Exemplo	**SS**	O próximo objetivo [da seleção]1_CS/TS é a [medalha de ouro nos Jogos Pan-Americanos do Rio]2_CS/ES. (S4_C28)
	SD	[Seu]1_CS/TS próximo objetivo é [os Jogos Pan-Americanos do Rio]2_CS/ES (S13_C28)
Comentário	A CS ocorre pela junção de material advindo da SD alinhada por especificação e de SD não-alinhadas por especificação. Por exemplo, o trecho da SS [medalha de ouro] é oriundo da S3 do D2 ("*Agora, a meta é a medalha de ouro dos Jogos Pan-Americanos, que estão sendo disputados no Rio de Janeiro.*"), cujo alinhamento à SS foi anotado como "neutro".	
Operação	**Transformação sintática**	
Etiqueta	TS	
Definição	Operação pela qual uma SS (ou um trecho dela) expressa conteúdo especificado resultante da transformação sintática da(s) SD(s) (ou de parte dela(s)).	
Exemplo	**SS**	O próximo objetivo [da seleção]1_CS/TS é a [medalha de ouro nos Jogos Pan-Americanos do Rio]2_CS/ES. (S4_C28)
	SD	[Seu]1_CS/TS próximo objetivo é [os Jogos Pan-Americanos do Rio]2_CS/ES (S13_C28)
Comentário	No exemplo, "aquele que possui o objetivo" é recuperado na SD pelo pronome anafórico "seu", ao passo que, na SS, essa informação está explícita por meio do sintagma preposicional "da seleção", que complementa o nome "objetivo".	

Operação	**Expansão sintagmática**	
Etiqueta	ES	
Definição	Operação responsável pela especificação, na SS, de uma palavra ou sintagma de uma SD pelo acréscimo de outros elementos; o resultado da expansão constitui-se em um sintagma mais complexo que o original.	
Exemplo	**SS**	O próximo objetivo [da seleção]1_CS/TS é a [medalha de ouro nos Jogos Pan-Americanos do Rio]2_CS/ES. (S4_C28)
	SD	[Seu]1_CS/TS próximo objetivo é os [Jogos Pan-Americanos do Rio]2_CS/ES (S13_C28)
Comentário	No exemplo, o sintagma original "Jogos Pan-Americanos do Rio" foi especificado para "medalha de ouro nos Jogos Pan-Americanos do Rio".	
Operação	**Expansão lexical**	
Etiqueta	EL	
Definição	Operação que especifica, na SS, uma palavra de uma SD pelo acréscimo de um especificador.	
Exemplo	**SS**	A Casa Branca em nota divulgada pela imprensa local considerou o atentado o pior ataque a tiros em um [campus universitário]1_CS/EL da história dos [Estados Unidos]2_CS/EL. (S9_C18)
	SD	Pelosi, por sua vez, pediu um minuto de silêncio na Câmara e disse que "se trata do pior tiroteio em um [campus]1_CS/EL na história do [país]2_CS/EL". (S22_C18)
Comentário	A informação sobre o local foi especificada pela expansão da palavra "campus" para "campus universitário".	
Operação	**Substituição lexical**	
Etiqueta	SL	
Definição	Operação em que determinado conteúdo é especificado no SS pela troca de uma palavra (ou expressão multipalavra) mais genérica por uma mais específica.	
Exemplo	**SS**	A Casa Branca em nota divulgada pela imprensa local considerou o atentado o pior ataque a tiros em um [campus universitário]1_CS/EL da história dos [Estados Unidos]2_CS/EL. (S9_C18)
	SD	Pelosi, por sua vez, pediu um minuto de silêncio na Câmara e disse que "se trata do pior tiroteio em um [campus]1_CS/EL na história do [país]2_CS/EL". (S22_C18)
Comentário	No caso 2 do exemplo, a informação sobre o local em que o evento ocorreu foi especificada por meio da troca da palavra "país" pelo nome próprio do país, "Estados Unidos".	
Operação	**Substituição sintagmática**	
Etiqueta	SSi	
Definição	Operação em que determinado conteúdo é especificado na SS pela troca de um sintagma genérico, proveniente de uma SD, por um mais específico.	
Exemplo	**SS**	O Brasil [conseguiu um gol logo nos primeiros 4 minutos]1_CS/SSi do jogo, fazendo os argentinos apertarem o ataque no jogo, restando ao Brasil os contragolpes, chegando ao segundo gol, que foi um gol contra.
	SD	É verdade que o Brasil deu sorte de [conseguir um gol logo no início da partida.]1_CS/SSi
Comentário	No exemplo, o sintagma "início da partida" foi especificado para "nos primeiros 4 minutos".	
Operação	**Substituição sentencial**	
Etiqueta	SSe	
Definição	Operação em que determinado conteúdo é especificado na SS pela troca de uma palavra ou sintagma genérico, de uma SD, por uma oração ou período composto.	
Exemplo	**SS**	Segundo Lula, [o mundo precisa de uma nova matriz energética, e o etanol pode ser mais que uma alternativa de energia limpa]1_CS/SSe.
	SD	Precisamos avaliar o caminho percorrido e estabelecer [novas metas]1_CS/SSe
Comentário	No exemplo, o sintagma "novas metas" foi especificado por uma oração coordenada.	

Operação	Substituição numérica	
Etiqueta	SN	
Definição	Operação em que determinado conteúdo é especificado na SS pela troca de uma palavra ou expressão, proveniente de uma SD, por um dado numérico.	
Exemplo	**SS**	[A pista principal]1_CS/SSi do [Aeroporto Internacional de São Paulo (Cumbica), em Guarulhos]2_CS/EL será totalmente reformada a partir de março de [2008]3_CS/SN, segundo informações do Ministério da Defesa.
	SD	O Ministério da Defesa anunciou nesta segunda-feira (6) que em março do [ano que vem]3_CS/SN [uma das pistas]1_CS/SSi do [Aeroporto de Guarulhos]2_CS/EL será fechada para reformas de seu trecho central.
Comentário	No exemplo, a expressão "ano que vem" foi especificada por "2008".	

Quadro 5: Caracterização das operações de reescrita dos casos de especificação do CSTNews.

Na Tabela 2, apresenta-se a distribuição quantitativa das operações nos alinhamentos.

Tabela 2: Estatística das operações de "recorta e cola" nos alinhamentos com especificação.

Operação	Quantidade absoluta	Porcentagem
Combinação sentencial	42	50%
Expansão sintagmática	11	13,09%
Substituição sentencial	10	11,90%
Substituição lexical	10	10,70%
Substituição sintagmática	6	9,50%
Expansão lexical	3	2,38%
Transformação sintática	1	1,20%
Substituição numérica	1	1,20%
TOTAL	**84**	**100%**

Com base nos dados quantitativos da Tabela 2, apresentam-se algumas observações quanto às operações de "recorta e cola" referentes à especificação no CSNews:

- A operação "combinação sentencial" é a mais frequente nos casos de especificação do CSTNews, com 50% das ocorrências, tendo sido observada em todas as sentenças dos sumários com conteúdo especificado (cf. Quadro 8);
- As operações "expansão sintagmática", "substituição sentencial", "substituição lexical" e "substituição sintagmática" foram um bloco intermediário, cujas frequências de ocorrência são bastante próximas, ou seja, 13,09%, 13,09%, 11,9% e 7,14% respectivamente;
- As operações "expansão lexical", "transformação sintática" e "substituição numérica" formam o bloco das menos frequentes; essas operações apresentam frequências também bastante próximas, ou seja, 2,3%, 1,2% e 1,2%, respectivamente;
- A grande maioria dos casos de "combinação sentencial" caracteriza-se pelo fato de a SS estar alinhada a apenas uma SD por especificação (cf. Tabela 1); assim, o trecho especificado presente no sumário veicula material linguístico proveniente de outra S da coleção, à qual a SS fora alinhada por causa de outro processo de fusão ("generalização", "inferência" ou "neutro").

No Quadro 6, apresenta-se essa distribuição em função dos *clusters*.

Corpus	Operações de reescrita							
Cluster	Combinação sentencial	Transformação sintática	Expansão sintagmática	Expansão lexical	Substituição lexical	Substituição sintagmática	Substituição sentencial	Substituição numérica
C2	1						1	
C4	1							
C9	1		1					
C12	1						1	
C18	2			1	1			
C21	3			1		1		1
C22	2		1				1	
C25	2		2					
C25	1					1		
C26	2					1	1	
C26	1		1					
C26	1				1			
C27	1						1	
C27	1						1	
C27	2				1		1	
C28	1	1	1					
C29	1				1			
C32	1					1		
C33	1						1	
C33	1						1	
C34	1		1					
C34	1						1	
C35	3				2	1		
C35	1				1			
C36	2				2			
C36	2		2					
C37	1					1		
C39	1							
C41	1			1				
C45	2		2					
	42	**1**	**11**	**3**	**9**	**6**	**10**	1

Quadro 6: Tabulação das operações de "recorta e cola" dos casos de especificação do CSTNews.

4.4. Análise das operações de "recorta e cola" envolvidas na especificação

Com base nos dados do Quadro 6, analisaram-se as ocorrências de cada uma das operações identificadas no *corpus* com o objetivo de identificar regularidade nos processos de especificação. Tais regularidades subsidiariam a proposição das estratégias descritas na próxima subseção. Para a descrição dessas análises, a seguir, parte-se das operações menos frequentes para as mais frequentes do *corpus.*

4.4.1. Especificação por Substituição numérica

O único caso de especificação que ocorreu por meio de uma substituição numérica, como definido, caracteriza-se pela troca da expressão "ano que vem" pelo ano específico, no caso, "2008".

4.4.2. Especificação por Transformação sintática

Assim como a substituição numérica, apenas um caso de especificação por transformação sintática foi identificado no *corpus* CSTNews. Nesse caso, o trecho "seu próximo objetivo" foi especificado para "o próximo objetivo da seleção". Aqui, entende-se "aquele que possui o objetivo" é recuperado na SD pelo pronome anafórico "seu", ao passo que, na SS, essa informação está explícita por meio do sintagma preposicional "da seleção", que complementa o nome "objetivo".

4.4.3. Especificação por Expansão lexical

Com base nos 3 casos de especificação por expansão lexical, foi possível identificar dois padrões. Um deles se caracteriza pela especificação de um item lexical pelo acréscimo, a ele, de um especificador, que pode ser um adjetivo (p.ex.: de "em um campus" para "em um campus universitário"). A outra regularidade no processo de especificação se caracteriza pela troca de um nome próprio mais popular de uma entidade (p.ex.: aeroporto) pelo nome completo da mesma, que é mais especifico que o popular, pois expressa informação que não está presente no nome popular. No caso da expansão de "Aeroporto de Guarulhos" para "Aeroporto Internacional de São Paulo [, em Guarulhos]", vê-se que as informações sobre o nível do aeroporto ("internacional") e o estado no qual está localizado ("São Paulo") estão dispostas na expressão especificada e não na expressão original. A outra se caracteriza pela adição, ao nome de uma entidade (pessoa), de sua profissão e sobrenome (p.ex.: de "Thiago" para "o nadador Thiago Pereira").

4.4.4. Especificação por Substituição sintagmática

Os casos de especificação por "substituição sintagmática" são bastante diferentes entre si. Assim, buscou-se delimitar o processo de especificação em função de cada caso, sempre com o objetivo de generalizar a substituição sintagmática para que as estratégias correspondentes também sejam aplicáveis a casos similares e não somente ao caso ocorrido no *corpus*. Dentre os 6 casos, foi possível delimitar a especificação em 5 deles[7]. O primeiro diz respeito à substituição do sintagma "uma das pistas" para "a pista principal". Do ponto de vista semântico, vê-se aqui que a expressão especificada ("a pista principal") veicula uma espécie de hipônimo da expressão original, já que "a pista principal" pode ser vista como um dos tipos de pista do aeroporto. Do ponto de vista na materialidade linguística, o sintagma original genérico "um(a) de(as) *x*, foi substituído por "a *x y*". No caso, os nomes podem ocorrer na posição de *x* e os adjetivos, na posição de *y*. No segundo caso, o sintagma preposicional (SPrep) "no início", em "[conseguir um gol logo] no início", foi substituído por outro SPrep, que indica o tempo específico em que o gol ocorreu, tendo-se "[conseguir um gol logo] nos primeiros 4 minutos". No terceiro caso, o sintagma nominal "o governo dos Estados Unidos" foi substituído por "a polícia norte-americana". Aqui, pode-se estabelecer que a especificação, do ponto de vista

[7] No caso em que o SPrep "ao sul da Jamaica", em "[o furacão] passou ao sul da Jamaica", foi transposto para "pela costa da Jamaica", não foi possível identificar, de fato, uma especificação.

semântico, ocorreu por meio da meronímia, já que a polícia pode ser vista como parte que compõe o governo. O quarto caso caracteriza-se pela troca do sintagma "grandes danos materiais" por "casas e viadutos destruídos". Novamente, pode-se pensar em uma substituição por hiponímia, já que "a destruição de casas e viadutos" pode ser concebida como um tipo de dano material (hiperônimo). O quinto caso é um exemplo em que a especificação ocorreu por meio do acréscimo de um aposto, já que, ao sintagma nominal "a seleção brasileira", foi adicionado "sob direção de Dunga", resultando na expressão especificada "a seleção brasileira, sob direção de Dunga".

4.4.5. Especificação por Substituição lexical

Dos 9 casos de substituição lexical, foi possível delimitar o processo em 8 deles. Desses 8, 2 deles se caracterizam pela especificação ocorrer por meio do emprego de hipônimos clássicos, ou seja, expressões que indicam "tipos de *x*". No *corpus*, os 2 casos de especificação por hiponímia foram: (i) "armas" > "revólveres", já que estes são um tipo de arma, e (ii) "abuso sexual" > "pedofilia". Em outros 6 casos, as palavras "país", "ilhas", "parlamentar", "colombiano" e "adversário" foram substituídas pelos nomes próprios das entidades, a saber: (i) "país" > "Estados Unidos"; (ii) "país" > "Brasil", (iii) "ilhas" > "Ilhas Cayman", (iv) "parlamentar" > "ACM", (v) "colombiano" > "Abadia" e (vi) "adversário" > "Vieri.

4.4.6. Especificação por Substituição sentencial

Dos 10 casos de especificação por "substituição sentencial", foi possível delimitar o processo em 3 deles. Desses, 1 ocorreu em um *cluster* da categoria esporte. Trata-se da substituição da palavra "*golaço*" por "*x chutou no ângulo*". O segundo caso delimitado de "substituição sentencial", ocorreu em um *cluster* da categoria política. Em especial, a expressão "*variação na lista espontânea*" foi substituída por uma sentença que especifica a variação das intenções de voto para cada candidato da lista (no caso, "*Lula sobe de 27% para 31%, Geraldo de 4% a 14% e Heloisa de 1% a 6%*"). Em outro caso, a especificação se deu pelo emprego de uma sentença que expressa a causa ("*um nevoeiro que cobria a região sul de São Paulo*") de certo efeito ("*problemas em Congonhas*").

4.4.7. Especificação por Expansão sintagmática

Os casos de especificação por "expansão sintagmática" são bastante variados, caracterizando-se por processos semânticos distintos. Dos 11 casos, foi possível delimitar o processo de expansão sintagmática em 7 deles, a saber:

1. uma lista inicial 2 de crimes contra a administração pública foi especificada pelo acréscimo de outros tipos de crimes administrativos ("*desvio de recursos públicos e influência indevida*" → "*desvio de recursos públicos, corrupção, prevaricação, concussão, peculato, extorsão, lavagem de dinheiro e venda de sentenças judiciais*");

2. a conquista de um campeonato esportivo foi especificada pelo acréscimo do nome da competição e do placar do jogo que decidiu o campeonato ("[...] *abocanhou seu oitavo título da competição continental.*" → "[...] *conquistou o oitavo título da Copa América, goleando a Argentina por 3 a 0.*");
3. a próxima meta de um time é conquistar determinado campeonato foi especificada pela informação de que o objetivo é a conquista da medalha de ouro no referido campeonato ("*O próximo objetivo é os Jogos Pan-Americanos do Rio.*" → "*O próximo objetivo é a medalha de ouro nos Jogos Pan-Americanos do Rio.*");
4. a informação de que *x* intensificou a fiscalização foi especificada pelo acréscimo do alvo específica da fiscalização ("*A polícia Federal intensificou a fiscalização* e o resultado [...]" → "*A polícia Federal intensificou a fiscalização sobre as declarações das pessoas físicas* [...]");
5. determinada operação investigativa foi especificada pelo acréscimo de (i) o nome da operação, (ii) o órgão responsável e (iii) o seu objetivo (p.ex.: "[...] *a operação havia capturado* [...]" → ["*A Operação Farrapos, da Polícia Federal, com o objetivo de desarticular uma quadrilha internacional de tráfico de drogas* [...]");
6. a referência a alguém com base na sua profissão é especificada pelo acréscimo do nome completo ("*Segundo a polícia, o garçom é fugitivo* [...]" → "[...] *é o garçom Wagner do Nascimento Marinho* [...]);
7. a informação de que determinado local fica em determinada região do estado foi especificada pela informação de que o local fica a *x* km da capital ("[...] *penitenciária de Valparaíso, no interior paulista.*" → "[...] *penitenciária de Valparaíso, a 580Km da capital.*").

4.4.8. Especificação por Combinação sentencial

Tendo em vista a frequência elevada de ocorrência da operação de "combinação sentencial" no *subcorpus*, observa-se que, de fato, os humanos combinam informação (e material linguístico) de mais de uma sentença dos textos-fonte para especificar o conteúdo a compor o sumário. Com base no Quadro 8, aliás, pode-se dizer que a combinação sentencial é uma operação necessária às demais, já que sempre uma operação ocorre, ocorre também a combinação sentencial. Para ilustrar como a combinação sentencial ocorre, considera-se a sentença do sumário: "*O último gol veio depois de um chute fraco de Kaká e um frango do goleiro Vieri*".

Esta foi alinhada a 2 sentenças-fonte por especificação:

(i) "*Na volta da Seleção Brasileira ao Maracanã, os jogadores não decepcionaram e o Brasil goleou o Equador por 5 a 0, com direito a golaço, jogada bonita, show de dribles e frango do goleiro adversário*" (S1_D4_C27);

(ii) "*Os mais de 80 mil torcedores que lotaram o estádio fizeram uma linda festa e aplaudiram muito os craques, principalmente Kaká, autor de dois gols, Ronaldinho, que fez um, e Robinho, que deu passe para outro, após lindíssima jogada.*" (S2_D4_C27).

Os trechos das sentenças-fonte que supostamente serviram de base para identificação da especificação foram [5 a 0], em S1, e [dois gols], em S2. Diz-se supostamente porque é muito difícil identificar com certeza os trechos nos quais os humanos de fato se pautaram para especificar o conteúdo. Considerando os trechos em questão, entende-se a sentença do sumário específica a jogada que resultou no último dos 5 gols da partida, o qual foi marcado por Kaká.

Para expressar no sumário essa informação especificada, pode-se dizer que os humanos se basearam em 2 sentenças-fonte em especial, as quais estão descritas abaixo. Nelas, os trechos sublinhados parecem ter sido os mais relevantes para o processo de especificação realizado pelo humano, que resultou na sentença do sumário (isto é, "*O último gol veio depois de um chute fraco de Kaká e um frango do goleiro Vieri*").

(i) "*Aos 39, Kaká chutou fraco e o camisa 1 tomou um frango histórico.*" (S33_D4)
(ii) "*Quando a torcida já cantava e gritava olé, Kaká ainda teve vontade para fazer mais um, só que o quinto gol brasileiro foi muito mais mérito do arqueiro Viteri, que não segurou um chute fraco e no meio do gol dado pelo meia brasileiro.*" (S18_D2)

Assim concebida, a operação de "combinação sentencial" é bastante complexa e, por isso, não se pode identificar padrões ou regularidades, ao menos neste ponto da pesquisa, sobre a forma de combinação das informações para a especificação do conteúdo. Assim, não foram delimitadas estratégias/regras para os casos de "combinação sentencial".

5. Proposição de estratégias automáticas de "recorta e cola"

Uma vez delimitadas, as operações de reescrita revelam-se como estratégias humanas de especificação, as quais foram traduzidas em regras no formato lógico *se, então,* tratável por máquina.

Especificamente, essas regras codificam as condições (*se*) de ocorrência da especificação nos textos-fonte (ou seja, os padrões identificados nos trechos dos textos-fonte) e a generalização *(então*) a ser feita (ou seja, o padrão identificado no(s) sumário(s)). Assim, dada uma coleção nova a ser sumarizada, o sistema sabe que, se uma das condições ocorrer nos textos-fonte, há um processo de especificação a aplicar.

No Quadro 7, sistematizam-se as regras referentes a cada uma das operações de "recorta e cola". Nas regras, os operadores lógicos estão em negrito (**se, então**). O itálico indica a ocorrência de uma palavra ou expressão fixa na condição e/ou na ação da regra.

Por exemplo, na regra para "substituição numérica", a expressão da condição (**Se**) é fixa (no caso, *ano que vem*). Na ação (**Então**), as aspas indicam que se trata de

uma expressão genérica; assim, na Regra 1, a expressão fixa *ano que vem* pode ser especificada por qualquer expressão que indica ano.

Operação "recorta e cola"	**Estratégia/Regra**	**Qt.**
Substituição numérica	**Se** [*ano que vem*] **Então** especificar para [expressão que indica ano] (p.ex.: *2008*)	1
Transformação sintática	**Se** [SN (pronome possessivo + {adjetivo}[8] + nome1 ∧ SN (sujeito)] (p.ex.: *seu próximo objetivo*) **Então** especificar para [SN (artigo + {adjetivo} + nome1) + Sprep (*de* + nome2) ∧ SN (sujeito)] (p.ex.: *o próximo objetivo da seleção*; nome2 explicita o referente do pron. possessivo)	2
Expansão lexical	**Se** [SPrep (preposição + artigo + nome)] (p.ex. *em um campus*) **Então** especificar para [Sprep (preposição + artigo + nome + adjetivo)] (*[9] o adjetivo especifica o nome) (p.ex.: *em um campus universitário*)	3
	Se [SN (nome próprio de pessoa)] (p.ex.: *Thiago*) **Então** especificar para [SN (artigo + nome (=profissão) + nome próprio (=pessoa))] (p.ex.: *o nadador Thiago*)	4
	Se [SN (nome popular de entidade)] (p.ex.: *Aeroporto de Guarulhos*) **Então** especificar para SN (nome completo da entidade) (p.ex: *Aeroporto Internacional de São Paulo*)	5
Substituição sintagmática	**Se** [SN [*um(a) de(as,os)* x (=nome plural)]] (p.ex.: *uma das pistas*) **Então** especificar para [SN [*o(a)* x (=nome singular) y (=adjetivo singular)]] (p.ex.: *a pista principal*)	6
	Se [SPrep [*em(o) início* (do jogo)]] **Então** especificar para [SPrep [*aos* x (=numeral) *minutos* (do jogo)]]	7
	Se [SN [*o governo de* x (=nome de país)]] **Então** especificar para [SN [*a polícia de* x (=nome de país) ∨ *a polícia* y (=adjetivo pátrio)]]	8
	Se [SN (*danos materiais*)] **Então** especificar para [SN (*casas e viadutos destruídos*)]	9
	Se [SN [*seleção* y (=adjetivo pátrio)]] **Então** especificar para [*seleção* y (=adjetivo pátrio), *sob o comando de* w (=nome próprio/técnico)] ∨ [*seleção* y (=adjetivo pátrio), *comandada por* w (=nome próprio/técnico)]	10
Substituição lexical	**Se** [*ilhas*] **Então** especificar para [nome próprio] (das ilhas) (p.ex.: *Ilhas Cayman*)	11
	Se [*país*] **Então** especificar para [nome próprio] (do país) (p.ex.: *Brasil*)	12
	Se [*parlamentar*] **Então** especificar para [nome próprio] (do parlamentar) (p.ex.: *ACM*)	13
	Se [SN (*o(a,os,as)* nome gentílico)] (p.ex: *o colombiano*) **Então** especificar para [SN (nome próprio)] (da pessoa) (p.ex.: *Abadia*)	14
	Se [SN [x (=nome; posição no time) *adversário*] (p.ex.: *goleiro adversário*) **Então** especificar para [SN [x (=nome; posição no time) nome próprio]] (=pessoa) (p.ex.: *goleiro Vieri*)	15
Substituição sentencial	**Se** [*golaço*] **Então** especificar para S [x (=nome próprio) *chutou no ângulo*]	16

[8] O símbolo { } indica que o elemento por ele delimitado é facultativo.
[9] O símbolo (*) indica um comentário a respeito da estratégia/regra.

	Se [SN (*variação na lista espontânea*)] **Então** especificar para Lista de ao menos 2 orações coordenadas no formato [x (nome próprio) *sobe de* y(=número) *% para* w(=número)]	17
	Se [*decorrente/ devido a/por causa dos problemas em* nome de aeroporto] (p.ex.: *por causa de problema em Congonhas*) **Então** [*decorrente/ devido a/por causa de*(o,os,a,as) SN (núcleo/nome que indica condição climática)] (p.ex.: *por causa de um nevoeiro em São Paulo*)	18
Expansão sintagmática	**Se** [SN (*desvio de recursos públicos e influência indevida*)] **Então** especificar para [SN (*desvio de recursos públicos, corrupção, prevaricação, concussão, peculato, extorsão, lavagem de dinheiro e venda de sentenças judiciais*)]	19
	Se [S (*x* conquistar título em competição)] **Então** especificar para S [*x* conquistar título na competição *y* (=nome próprio) com o placar de *w*"] (o símbolo S indica "sentença")	20
	Se [S (*o próximo objetivo é* o(os,a,as) nome de competição esportiva)] **Então** especificar para [S (*o próximo objetivo é a medalha de ouro* em(o,os,a,as) nome de competição esportiva)]	21
	Se [*fiscalização*] **Então** especificar para [*fiscalização sobre/de* x (SN)]	22
	Se [SN [a *operação* (investigativa)]] **Então** especificar para [SN [*a operação* x (=nome próprio1), *de*(a,as,o,os) y (=nome próprio2), *com o* nome (=*objetivo, meta*) *de*]]	23
	Se [SN [artigo + nome (=profissão)]] (p.ex.: *o garçom*) **Então** especificar para [SN [artigo + nome (=profissão) + nome próprio]] (p.ex.: *o garçom Wagner do Nascimento Marinho*)	24
	Se [*no interior* + adjetivo (=gentílico) (p.ex.: *no interior paulista*) **Então** especificar para [*a* x (=número) *km da capital*)	25

Quadro 7: Formalização das estratégias de especificação em regras no formato lógico.

6. Considerações finais

Como resultados, destacam-se a identificação e a tipificação das 84 operações de "recorta e cola" dos 40 casos de especificação CSTNews e a geração do manual composto pela definição e exemplificação dos 8 diferentes tipos de operação de reescrita (cf. Quadro 6), o qual poderá ser utilizado em trabalhos futuros. Além desses, destaca-se a proposição das 25 estratégias formais de especificação, de acordo com os casos do *corpus*. Essas estratégias, no formato de regras lógicas, poderão subsidiar a geração automática de sumários com especificação.

Sobre as estratégias, em especial, salienta-se que, apesar do esforço em elaborar estratégias genéricas, algumas das regras são ainda bastante pontuais. Algumas delas, aliás, são até mesmo lexicalizadas, ou seja, dependentes da ocorrência de uma palavra ou expressão específica. Esse é caso, por exemplo, da Regra 1, que se aplica somente no caso da ocorrência da expressão "*ano que vem*". Outras, por se baseiam na informação de classe de palavras, são mais genéricas (p.ex.: Regra 2).

Além disso, destaca-se que, para a aplicação computacional de estratégias como a formalizada pela Regra 1, os sistemas de PLN (em especial, os sumarizadores automáticos) terão de realizar uma análise lexical dos textos-fonte que compreenderá: (i) a identificação da expressão "ano que vem" nos textos-fonte e (ii)

identificação das possíveis "expressões que indicam ano" nos próprios textos-fonte, pois estas poderão ser utilizadas para especificar, no sumário, a informação genérica original. A identificação de possíveis expressões especificadas nos textos-fonte é fundamental, pois a alta frequência da operação "combinação sentencial" revela que a especificação é um processo em que os humanos utilizam material advindo dos próprios textos-fonte para compor o sumário.

Assim, para aplicar a Regra 1, por exemplo, os sistemas precisarão saber quais as expressões em português que indicam ano, o que comumente é feito pelo acesso a um recurso linguístico (por exemplo, um léxico) que armazenada essa informação. Somente acessando um recurso desse tipo que a máquina conseguirá reconhecer a informação original genérica e as expressões que potencialmente poderão especificar essa informação no sumário. No caso das regras específicas do Quadro 8, tem-se que construir tais recursos, pois as informações necessárias à aplicação das regras não estão descritas em recursos linguísticos disponíveis.

O mesmo pode ser dito para o reconhecimento das designações genéricas (p.ex.: "Aeroporto de Guarulhos") e especificadas ("Aeroporto Internacional de São Paulo") de uma mesma entidade nomeada[10]. Nesse caso, os sistemas também necessitarão reconhecer e classificar (isto é, identificar se a entidade é uma *pessoa*, um *local*, etc.) as entidades nomeadas, o que comumente é feita por uma ferramenta denominada "reconhecedor de entidades nomeadas" (em inglês, *named entity recogniton*) (NER). Para o português, tem-se o reconhecedor Rembrandt (CARDOSO, 2008). Como muitas das entidades que ocorrem nos *corpora* jornalísticos como o CSTNews são dependentes de domínio, as bases de conhecimento do Rembrandt talvez tenham de ser estendidas pelo acréscimo de tal conhecimento. Dessa forma, o NER poderá, por exemplo, identificar as expressões ACM (isto é, Antônio Carlos Magalhães) e Abadia como entidades nomeadas e correlacionar as diferentes designações de uma mesma entidade (p.ex.: "Aeroporto de Guarulhos" e "Aeroporto Internacional de São Paulo"), além de classificar todas as entidades adequadamente (p.ex.: ACM (pessoa), Abadia (pessoa), Aeroporto de Guarulhos (local) e Aeroporto Internacional de São Paulo (local)).

Ademais, a aplicação de regras como a Regra 9 pode requerer a modelagem de certos conhecimentos de mundo (p.ex.: "destruição de casas e viadutos" é um tipo de "dano material"), os quais, por serem de domínio específico, não costumam constar nos recursos linguísticos do português.

Por fim, salienta-se que, por meio deste trabalho, sabe-se mais hoje sobre as operações de reescrita envolvidas no processo de especificação no cenário multidocumento do que antes do início do projeto. Como trabalho futuro, pretende-se avaliar a pertinência das estratégias/regras ora propostas.

Agradecimentos

Os autores agradecem à FAPESP pelo apoio financeiro.

Referências bibliográficas

AGOSTINI, V.; CAMARGO, R. T.; PARDO, T. A. S.; DI-FELIPPO, A. Alinhamento manual de textos e sumários em um corpus jornalístico

[10] São consideradas "entidade nomeadas" (ENs) quaisquer nomes próprios como locais, acontecimentos, nomes de pessoas, etc.

multidocumento. In: ENCONTRO DE LINGUÍSTICA DE CORPUS - ELC, 11, 2012, São Carlos/SP. Proceedings… São Carlos, 2012, p. 1-5.

AGOSTINI, V.; CAMARGO, R. T.; PARDO, T. A. S.; DI-FELIPPO, A. Manual Typification of Source Texts and Multi-document Summaries Alignments. São Carlos, 2013.

CAMARGO, R.T.; DI-FELIPPO, A.; PARDO, T.A.S. Em direção à caracterização de sumários humanos multidocumento. In: WORKSHOP ON PORTUGUESE DESCRIPTION, 2, 2011, Cuiabá. Proceedings... Cuiabá/MT/Brazil, 2011, p. 47-54. October 26, 2011.

CAMARGO, R.T. Investigação de estratégias de sumarização humana multidocumento. 2013. 133f. Dissertação (Mestrado em Linguística) – Departamento de Letras, Universidade Federal de São Carlos, São Carlos, 2013.

CARDOSO, N. Rembrandt - Reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. In MOTA, C.; SANTOS, D. (Eds.). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca. 2008

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

CASTRO JORGE, M.L.R.; PARDO, T.A.S. Experiments with CST-based Multi-document summarization. In: ACL WORKSHOP TEXTGRAPHS-5: GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING, 2010, Uppsala/Sweden. Proceedings…Uppsala, 2010, p. 74-82.

CREMMINS, E.T. The art of abstracting. Arlington, Virginia: Information Resources Press, 1996.

ENDRES-NIGGEMEYER, B. Summarization Information. Berlin: Springer, 1998.

JING, H.; MACKEOWN, K.R. The decomposition of human-written summary sentence. In: INTERNATIONAL ACM SIGIR, 22, 1999, New York. Proceedings… New York, 1999. p. 129-136.

_____. Cut and paste based text summarization. In: NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS CONFERENCE, 1., 2000, San Francisco, California. Proceedings…. San Francisco, 2000, p. 178-185.

JURAFSKY, D; MARTIN, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Prentice Hall, 2007. p. 1024.

KUMAR, Y.J.; SALIM, N. Automatic multi-document summarization approaches. Journal of Computer. Science, 8, 2012, p. 133-140.

LAGE, N. A reportagem: teoria e técnica de entrevista e pesquisa jornalística. Rio de Janeiro: Record, 2004.

_____. Estrutura da Notícia. 5ª ed. São Paulo: Ática, 2002.

LUHN, H. P. The automatic creation of literature abstracts. IBM Journal of Research, Riverton, v. 2, n. 2, p. 159-165, 1958.

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

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

NENKOVA, A. Understanding the process of multi-document summarization: content selection, rewrite and evaluation. PhD Thesis, Columbia University, January 2006.

RIBALDO, R.; AKABANE, A.T.; RINO, L.H.M.; PARDO, T.A.S. Graph-based methods for Multi-document Summarization: exploring relationship maps, complex networks and discourse information. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE (LNAI 7243), 10, 2012, Coimbra, Portugal. Proceedings... Coimbra, 2012, p. 260-271. April 17-20.

TEIXEIRA, H.M.L. O clipping de mídia impressa numa abordagem interdisciplinar sob os prismas da Ciência da Informação e da Comunicação Social: o jornal de recortes da Assembléia Legislativa de Minas Gerais. Perspectivas em Ciência da Informação. v. 6, n. 2, 2001. ISSN 1981-5344 (Online).