

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

***Análise Linguística da Operação de Generalização
na Sumarização Humana Multidocumento***

Marina Delege
Ariani Di Felippo

NILC-TR-14-01

Setembro, 2014

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

Resumo

Neste relatório, descreve-se um estudo sobre a generalização no CSTNews, *corpus* multidocumento jornalístico do português. A generalização é das várias operações de fusão *cross-document* identificadas no âmbito da sumarização humana. Por meio dela, expressa-se, no sumário, conteúdo mais genérico que o ocorrido nos textos-fonte. Essa investigação englobou a (i) descrição das operações de “recorta e cola” (p.ex.: substituição sintagmática, substituição lexical, redução sentencial, combinação sentencial, transformação sintática, lexicalização, substituição numérica, substituição causa-efeito, etc.) envolvidas no processo de generalização e a (ii) proposição de estratégias automáticas de generalização. Com isso, gerou-se conhecimento para subsidiar uma futura produção automática de *abstracts* (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 conta com o apoio financeiro da FAPESP (2013/12629-0).



1. Introdução

A sumarização multidocumento é uma atividade de produção textual que consiste na elaboração de uma versão condensada coerente e coesa, de dois ou mais textos-fonte, provenientes de diferentes fontes, que abordam um mesmo assunto. Os sumários multidocumento “informativos e genéricos”, mais tradicionais, veiculam a informação principal dos textos-fonte e são elaborados para uma audiência ampla (não específica). Como exemplo, tem-se um resumo que condensa as informações de uma coleção de notícias de fontes diferentes sobre um mesmo fato ou evento.

Segundo Crammins (1996) e Endres-Niggemeyer (1998), os humanos produzem um sumário em 3 etapas: (i) análise das características do(s) texto(s)-fonte, (ii) seleção do conteúdo relevante e (iii) produção do sumário. Para realizar cada um desses estágios, estratégias específicas de interpretação e/ou geração de texto são empregadas.

Para condensar o conteúdo selecionado, algumas operações de fusão de informação (do inglês, *cross-document fusion*) podem ser empregadas, como eliminação, união, intersecção, generalização/especificação, etc (MANI, 2001). Na produção do sumário, o conteúdo condensado pelas operações de fusão pode ser linguisticamente expresso com base em operações de *cut-and-paste* (de palavras, expressões, etc.) dos próprios textos-fonte, como redução sentencial, combinação sentencial, transformação sintática, paráfrase lexical e reordenação (JING, MACKEOWN, 1999; 2000).

Baseando-se no processo humano, tem-se desenvolvido, na Sumarização Automática Multidocumento (SAM), sumarizadores automáticos que geram sumários informativos e genéricos pela justaposição de sentenças selecionadas na íntegra dos textos-fonte (comumente, notícias jornalísticas) (p.ex.: CASTRO JORGE, 2010; RIBALDO et al., 2011). Nesse caso, os sumários são extrativos (ou extratos). Apesar de a boa coesão e coerência dos sumários multidocumento extrativos, os sumarizadores automáticos ainda não produzem *abstracts*, ou seja, sumários elaborados pela reescrita dos textos-fonte. Um dos motivos dessa limitação é a falta de conhecimento sobre as operações de (i) fusão de informação e (ii) *cut-and-paste* empregadas na SHM.

Assim, investigou-se a generalização no CSTNews (CARDOSO et al., 2011), *corpus* multidocumento jornalístico do português, descrevendo as operações de “recorta e cola” envolvidas nesse processo e identificando operações recorrentes que podem subsidiar a produção automática de *abstracts*. Inserindo-se no projeto SUSTENTO¹ (FAPESP 2012/13246-5/ CNPq 483231/2012-6), que busca gerar subsídios para a SAM em português, este trabalho produziu conhecimento específico sobre a generalização e propiciou ao aluno o exercício do método científico no PLN.

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 generalização. Na Seção 5, apresentam-se as estratégias automáticas de generalização aqui propostas. Por fim na Seção 6, tecem-se considerações finais sobre as atividades ora relatadas e apontam-se trabalhos futuros.

¹ <http://www.nilc.icmc.usp.br/arianidf/sustento/>

2. Revisão da literatura

Os tópicos revisados foram: (i) conceitos básicos sobre SHM, (ii) as operações básicas de fusão *cross-document* e (iii) as operações de “recorta e cola”.

2.1. A sumarização humana multidocumento

A sumarização humana pode ser entendida como o processo de seleção de conteúdo de um ou mais texto-fonte para produzir uma versão mais curta do(s) mesmo(s) visando determinado usuário/tarefa. Há vários tipos de sumários humanos, por exemplo: resenhas de notícias jornalísticas, sinopses do movimento da bolsa de valores, sumários de textos novelísticos, extratos de livros científicos, resumos de previsões meteorológicas, etc.

Compreender como os humanos realizam a sumarização, no entanto, não é nada fácil devido à subjetividade da tarefa. Vários autores têm buscado compreender e sistematizar o modo como os humanos geram versões condensadas a partir de um ou mais documentos. De modo geral, reconhece-se que a sumarização humana envolve alguns subprocessos (CREMMINS, 1996; ENDRE-NIGGEMEYER, 1998), os quais estão descritos no Quadro 1. Contudo, Endres-Niggemeyer (1990) observa que o modo como eles realizam cada um desses passos ainda pode variar de acordo com o indivíduo, dificultando a modelagem do processo.

Etapas	Objetivo	Resultado
Exploração do documento	Identificar as características básicas do material: 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.	Construção do “esquema” (do inglês, <i>scheme</i>), isto é, de um conhecimento inicial sobre o tipo de documento e estrutura da informação.
Avaliação da relevância	Identificar no texto unidades textuais relevantes, representar o texto em nível discursivo e combinar os elementos da etapa anterior, associando-os aos elementos da etapa em questão.	Construção do “tema” (do inglês, <i>theme</i>), ou seja, de uma representação mental estruturada sobre o conteúdo do texto.
Produção do sumário	Transportar itens textuais relevantes do documento original ao sumário e reorganizá-los em uma nova estrutura, envolvendo operações de “recorta e cola” (do inglês, <i>cutting and pasting operations</i>)	Produção de um texto condensado a partir dos pontos mais importantes de um documento.

Quadro 1: Etapas humanas de sumarização.

A produção de um sumário é influenciada por vários fatores. Além do tamanho desejado do sumário, que determina o quanto a informação do texto-fonte deve ser condensada, a sumarização é influenciada por: (i) a audiência a que se destina o sumário, (ii) a função do sumário, (iii) o tipo do sumário, (iv) a quantidade de textos-fonte, etc. (MANI, 2001).

Com base na audiência, pode-se produzir um sumário genérico, que veicula a informação principal do texto-fonte, substituindo a leitura do mesmo, ou um sumário que veicula a informação de interesse para um tipo específico de usuário. Quanto à função, pode-se objetivar a produção de sumários informativos, isto é, contém as informações principais de um texto-fonte ao ponto de dispensar a leitura do original, indicativos, ou seja, não substitui o texto-fonte, apenas diz do que ele trata (p.ex.: índices de livros) ou críticos, os quais apresentam, além da informação principal do texto-fonte, avaliações sobre ele (p.ex.: resenhas). Quanto ao tipo, têm-se os sumários extrativos (ou extratos), ou seja, compostos por trechos inalterados dos textos-fonte, e os abstrativos (ou *abstracts*), isto é, construídos pela reescrita das informações dos textos-fonte.

Quanto à quantidade de textos-fonte, a sumarização pode ser monodocumento ou multidocumento. Na sumarização humana multidocumento (SHM), em especial, o humano produz um resumo a partir de mais de um texto, provenientes de fontes distintas, que abordam um mesmo assunto. Apesar de pouco intuitiva, a sumarização multidocumento é relativamente comum em alguns domínios (MANI, 2001).

No âmbito jornalístico, o “*clipping*” é um exemplo de sumarização multidocumento. Essa atividade é bastante antiga na área da comunicação e se caracteriza pela pesquisa e seleção contínua de notícias relacionadas a determinados assuntos, atendendo a um público direcionado (TEIXEIRA, 2001).

Os *clippings* de mídia impressa, enquanto produto, podem ser veiculados nas formas: (i) clássica, ou seja, por meio de um conjunto de recortes de notícias, reportagens, artigos, etc., (ii) sinopse, isto é, por meio de um texto que contempla as principais notícias de interesse do cliente, ou (iii) análise, ou seja, por meio de um texto que contém uma interpretação crítica das informações coletadas (TEIXEIRA, 2001).

No caso do *clipping* eletrônico (*e-clipping* ou *web clipping*), a atividade consiste em selecionar, coletar e organizar as informações veiculadas por diversas fontes da *web* a respeito de determinado tópico, pessoa, instituição etc. Os resultados da seleção podem ser organizados e veiculados não só nas mesmas formas do *clipping* de mídia impressa, 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. Atualmente, várias empresas têm oferecido o serviço de *clipping* eletrônico, como a Associação Brasileira das Empresas de Monitoramento de Informação (ABEMO)², o Grupo Info4³, o Armazém Digital⁴, entre outras.

Além dessas empresas prestadoras de serviços, inúmeras instituições e associações também oferecem *clippings* eletrônicos a seus associados e/ou consultentes, como a Universidade Federal de São Carlos (UFSCar)⁵, Universidade de São Paulo (USP)⁶, Associação dos Advogados de São Paulo (AASP)⁷, entre outras.

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>

⁶ <http://www.usp.br/agen/clip/pdf/>

⁷ <http://www.aasp.org.br/aasp/imprensa/clipping/index.asp>

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.

Assim, 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.

Da “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).

Para condensar o conteúdo selecionado, algumas operações básicas de fusão de informação (do inglês, *cross-document fusion*) podem ser empregadas, 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 *cut-and-paste* (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).

2.2 As operações de fusão *cross-document*

A produção de sumários abstrativos que, por resultarem da reescrita dos textos-fonte são mais próximos dos sumários humanos, requer conhecimento sobre como o conteúdo dos textos-fonte é efetivamente condensado para a composição do sumário. Assim, a automação da geração de abstracts multidocumento requer a investigação das operações básicas de fusão *cross-document*, como deleção, fusão (união ou intersecção), generalização ou especificação e inferência de informação (MANI, 2001).

A eliminação, também denominada deleção, é a operação de condensação em que certas informações contidas nos textos-fonte são removidas ou excluídas. A fusão por união é uma operação por meio da qual as informações distintas, provenientes de textos distintos, são preservadas no sumário, eliminando-se a redundância. A fusão por intersecção consiste em combinar, no sumário, informações que se repetem nos textos-fonte. 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. No contrário, tem-se a especificação.

A seguir, ilustram-se separadamente as operações de deleção, união, intersecção, generalização, especificação e inferência. Em (1), há três sentenças extraídas de textos distintos que abordam uma mesma notícia. Tais sentenças veiculam conteúdo similar, ou seja, elas são redundantes entre si. Com base nessas três sentenças, ilustram-se as operações de união e intersecção que resultam nas sentenças em (2) e (3).

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

A aeronave da TAM Airbus A320, voo JJ3054, partiu de Porto Alegre, às 17h16 com destino a Congonhas. (Texto-fonte 2)

Um Airbus A320 com capacidade para 170 passageiros partiu de Porto Alegre (RS) às 17h16 com destino a Congonha. (Texto-fonte 3)

Com base na operação de fusão do tipo união, pode-se gerar o sumário em (2), no qual as informações distintas, provenientes de cada uma das sentenças-fonte, são preservadas no sumário, o qual não apresenta 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.*

No Quadro 2, discrimina-se a origem das informações contidas no sumário com o intuito de evidenciar a união das informações distintas de cada fonte. As informações específicas de cada fonte estão assinaladas com o símbolo “X”.

Elementos do sumário	Texto 1	Texto 2	Texto 3
A aeronave da TAM		X	
Airbus A320 / partiu de Porto Alegre / às 17h16	*	*	*
voo JJ3045	*	*	
com capacidade para 170 passageiros / (RS)			X
da terça-feira / chegou a São Paulo às 18h45	X		
com destino a Congonhas		*	*

Quadro 2: A origem das informações do sumário gerado pela operação de união

A partir das mesmas sentenças-fonte, é possível gerar o sumário em (3) com base na operação de fusão do tipo intersecção. Nesse sumário, somente as informações que mais se repetem nos textos-fonte (ou seja, redundantes) estão combinadas.

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

Para a condensação do conteúdo dos textos-fonte com base na intersecção, observa-se que as informações específicas de cada texto sofreram deleção, a saber: (i) “companhia

de voo”, do texto 2, (ii) “capacidade do avião” e “estado de onde o voo partiu”, do texto 3, e (iii) “dia da semana em que o voo partiu” e “horário de chegada do voo”, do texto 1.

A especificação é outra operação de fusão *cross-document*. Por meio dela, expressa-se conteúdo mais específico que o presente nos textos-fonte. Assim, a partir das mesmas sentenças do Quadro 1, pode-se gear por meio de inferências, por exemplo, uma sentença-sumário com informações mais específicas que as veiculadas nas sentenças-fonte, como a ilustrada em (4). Aliás, a especificação, assim como a generalização, pressupõe algum tipo de inferência.

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

Observa-se em (4) que o sumário apresenta duas informações mais específicas que as das sentenças que lhe deram origem. Com base principalmente em conhecimento de mundo, o produtor do sumário explicitou que “A320” indica um “modelo” de Airbus e que a origem do voo foi o aeroporto Salgado Filho.

Por fim, a generalização, foco deste projeto, é mais uma operação de condensação comumente utilizada para reduzir o conteúdo dos textos-fonte e produzir um sumário a eles correspondente. Por meio da generalização, o conteúdo dos textos de origem é expresso de forma mais abstrata ou genérica. Para tanto, o produtor do sumário também realiza um processo de inferência. Com base nas mesmas sentenças de (1), pode-se produzir uma sentença-sumário como a ilustrada em (5).

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

Observa-se em (5) ao menos dois casos de generalização. Um deles diz respeito ao fato de que o objeto (“aeronave”) envolvido no fato foi descrito sem as informações sobre seu modelo, número, capacidade de carga e companhia, as quais sofreram deleção. O outro caso refere-se à generalização do 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.

2.3. 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 *cut-and-paste* (de palavras, expressões, etc.) dos próprios textos-fonte, como (JING, MACKEOWN, 1999; 2000):

- (i) redução sentencial: 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: processo pelo qual a sentença de um sumário multidocumento é formulada pela combinação de elementos dos textos-fonte;
- (iii) transformação sintática: 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); esse processo engloba outro, o de reordenação;

(iv) paráfrase lexical: processo que se caracteriza pela alteração de vocábulos dos textos-fonte por sinônimos ou expressões equivalentes.

Em (2), por exemplo, observa-se que 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.

Em (3), 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. Esses elementos têm granularidade diferente (sintagmas preposicionais e sigla).

Em (5), 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”.

Na sequência, descrevem-se a seleção do *corpus* e a descrição das operações de “recorta e cola” envolvidas na generalização.

3. Seleção do *corpus*

Para investigar a operação de generalização no cenário multidocumento, selecionou-se o CSTNews (CARDOSO et al., 2011), *corpus* multidocumento de referência em português.

O CSTNews é composto por 50 coleções de textos jornalísticos, sendo que cada coleção versa sobre um mesmo assunto. Os textos são do gênero “notícias jornalísticas”, cujas propriedades são (i) documentar as experiências humanas vividas (domínio social) e (ii) representar pelo discurso as experiências vividas, situadas no tempo (capacidade da linguagem) (BARBOSA, 2001; DOLZ; SCHNEWLY, 2004; LAGE, 2002, 2004).

Cada coleção contém basicamente: (i) 2 ou 3 textos sobre um mesmo assunto, compilados de diferentes fontes jornalísticas e com tamanhos similares, (ii) sumários humanos mono e multidocumento, (iii) sumários automáticos multidocumento, e (v) diversas anotações linguísticas.

Os textos-fonte 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” (10 coleções), “mundo” (14), “dinheiro” (1), “política” (10), “ciência” (1) e “cotidiano” (14).

Quanto aos sumários multidocumento, ressalta-se que estes são informativos e genéricos e foram construídos manualmente de forma abstrativa, ou seja, por meio de operações de “recorta e cola” 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 se 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 1, 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.

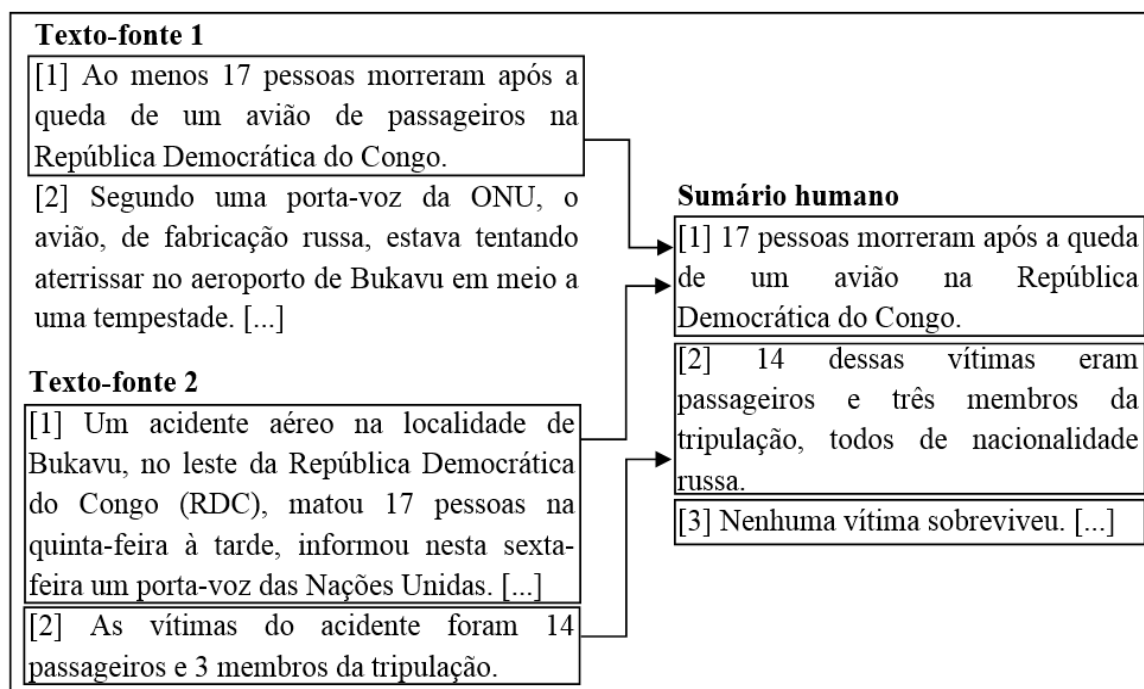


Figura 1: Exemplo de alinhamento sumário/textos-fonte do CSTNews.

O alinhamento entre o sumário e seus respectivos textos-fonte no interior de cada coleção foi feito em função de duas diretrizes centrais. A primeira delas diz respeito ao nível dos segmentos a serem alinhados e a segunda refere-se ao critério para a identificação das correspondências.

Quanto ao nível dos segmentos textuais, Agostini et al. (2012) optaram pelo sentencial, posto que as sentenças são unidades de informação bem delimitadas. Sobre o critério de alinhamento propriamente dito, ressalta-se que as correspondências entre os sumários e seus respectivos textos-fonte foram identificadas com base na sobreposição de conteúdo, total ou parcial. Tendo em vista que os sumários do CSTNews são *abstracts*, o alinhamento das informações contidas nos sumários com as dos textos-fonte nem sempre foi simples, posto que houve reescrita do conteúdo dos textos-fonte transpostos para a versão condensada.

Ao se optar por um alinhamento baseado na sobreposição de conteúdo ou informação, o processo de indexação não se baseia na sobreposição de formas, ou seja, de unidades lexicais. Consequentemente, sentenças que continham conteúdo em comum, total ou parcial, com baixa sobreposição lexical, foram alinhadas.

Os alinhamentos do CSTNews também foram manualmente tipificados. No caso, dado um par de sentenças alinhadas, a conexão entre elas foi caracterizada em função de: (i) a sobreposição de material linguístico e (ii) a operação de condensação.

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 os rótulos: (i) especificação, (ii) generalização, (iii) inferência, ou (iv) neutro (ou seja, quando o conteúdo de uma sentença do documento-fonte (SD) foi transposto à sentença do sumário (SSum) por alguma operação que não as codificadas pelas outras etiquetas).

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. 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.

O alinhamento em (6), retirado do CSTNews, recebeu as etiquetas: parcial, generalização, toponímia. A “toponímia”, em especial, deve-se ao fato de que a SSum generaliza a informação sobre os “estados de origem dos policiais”, inserindo a expressão “de vários estados” e excluindo a lista dos nomes dos estados.

(6) Sumário

Mais de 300 policiais federais de vários estados participaram das buscas e prisões durante a operação.

Texto-fonte

A PF divulgou que mais de 300 policiais federais do Amazonas, Distrito Federal, Mato Grosso, Acre e Rondônia fazem parte das investigações da "Operação Dominó".

Quanto à generalização, ressalta-se que, ao total, 47 sentenças distintas dos sumários foram alinhadas aos seus respectivos textos-fonte, gerando um total de 80 alinhamentos, posto que uma mesma SSum pode se alinhar a mais de uma SDs.

Tendo em vista o total de 1007 alinhamentos do CSTNews, os 80 alinhamentos do tipo **generalização** equivalem a aproximadamente 8% das conexões dos sumários aos textos-fonte do referido *corpus* (AGOSTINI et al., 2012; CAMARGO, 2013)

Os 80 casos do CSTNews foram manualmente revisados e, desse processo, 10 alinhamentos foram excluídos da tarefa de identificação das operações de “recorta e cola” por não englobarem efetivamente generalização. Tais alinhamentos devem ser resultado de algum tipo de interpretação equivocada do conteúdo dos textos-fonte e sumários. Os alinhamentos excluídos estão no Quadro 3.

Assim, das 47 SSum alinhadas, restaram 40 e, dos 80 alinhamentos, apenas 70 foram efetivamente alvo desta investigação.

Ressalta-se que, entre os 70 alinhamentos restantes, alguns deles se caracterizam pela conexão de uma SSum a apenas uma SD (ou seja, alinhamento 1-1) e outras se caracterizam pela conexão de uma SSum a mais de uma SD (ou seja, alinhamento 1-*n*). A

distribuição os diferentes tipos de alinhamento com generalização do CSTNews estão descritos na Tabela 1.

Caso	Ssum	SD
1	A Casa Branca em nota divulgada pela imprensa local considerou o atentado o pior ataque a tiros um campus universitário da história dos Estados Unidos. (S9)	Pelosi, por sua vez, pediu um minuto de silêncio na Câmara e disse que "se trata do pior tiroteio em um campus na história do país" (S22_D2)
2	A medalha de prata ficou com a americana April Steiner com 4m40 e a de bronze com a cubana Yarisley Silva com 4m30. (S2)	A atleta norte-americana provou que a disputa seria acirrada, já que bateu a marca de 4,40m, mesmo com as vaias da torcida local. (S11_D2)
3	A pressão argentina continuou no segundo tempo, mas o Brasil fechou a goleada com um gol marcado pela sua dupla de volantes. (S8)	O atacante, que na Copa América marcou apenas uma vez, deu um passe preciso para Daniel Alves chutar cruzado de dentro da área e vazar mais uma vez o goleiro Abbondanzieri. (S34_D1)
		Os argentinos, com um time repleto de craques, favoritos ao título e com uma campanha irrepreensível até o momento, pareciam não acreditar no que viam. (S17_D4)
		Vagner Love fez lindo passe para Daniel Alves, que chutou forte e cruzado no fundo das redes de Abbodanzieri. (S24_D5)
4	Antes de chegar à Jamaica, Dean matou ao menos nove pessoas nas ilhas de Santa Lúcia, Dominica, República Dominicana e Haiti, no Caribe. (S2)	No Haiti, quatro pessoas morreram, segundo as Nações Unidas, elevando para nove o número de mortos desde a chegada do Dean ao Caribe, no primeiro furacão da temporada de 2007, que deve ser muito ativa. (S14_D4)
		Os brasileiros que estão em Cancún, um dos pontos turísticos mais visitados do México, estão se preparando para a chegada do furacão Dean, que já matou nove pessoas na região do Caribe. (S1_D5)
5	Na próxima semana, o Conselho de Ética vai decidir sobre a unificação dos processos contra o presidente da Casa, Renan Calheiros. (S1)	Ficou fatiado nesse monte de processos porque o conselho assim o quis", criticou Nery. (S15_D1)
6	O próximo confronto será contra os rivais mais perigosos, a seleção de Cuba. (S9)	Quantas vezes não fui vaiado porque os torcedores queriam ver o William jogar?", perguntou o técnico. (S12_D1)
7	O líder do governo no Senado, Romero Jucá (PMDB), admitiu que é possível. (S6)	Dissemos que a proposta é insuficiente, mas sentimos que o governo ainda tem gordura para queimar - afirmou Virgílio. (S5_D3)

Quadro 3: Casos de generalização do CSTNews desconsiderados.

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

	Tipos de alinhamento					
	1-1	1-2	1-3	1-5	1-6	1-12
Número de alinhamentos	30	3	4	1	1	1

Com base na Tabela 1, pode-se dizer que cada uma das 30 sentenças dos sumários nos alinhamentos do tipo simples (1-1) foi elaborada a partir do conteúdo e do material linguístico de apenas uma sentença, ao passo que cada SSum nos alinhamentos complexos (1-2, 1-3, 1-5, 1-6 e 1-12), foi construído a partir de conteúdo e material de linguístico de mais de uma SD. Na sequência, descreve-se a tarefa de identificação das operações de “recorta e cola” envolvidas nos 70 casos de generalização selecionados.

4. Identificação das operações de “recorta e cola”

A identificação das operações de “recorta e cola” englobou 3 tarefas: (i) identificação, delimitação e indexação dos casos de generalização nos 70 alinhamentos; (ii) identificação, anotação e descrição das operações de “recorta e cola”, e (iii) organização e registro dos dados de análise.

4.1. Identificação, delimitação e indexação dos casos de generalização

Os 70 alinhamentos do CSTNews foram manualmente analisados de forma individual com o objetivo de identificar efetivamente os casos de generalização, pois uma SSum associada a uma única SD pode apresentar mais de um caso de generalização, assim como uma SSum alinhada a várias SDs pode conter apenas uma generalização.

Tal identificação englobou a delimitação e a indexação dos trechos generalizados das SSum e dos trechos de origem das 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*.

Em (7), 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 SSum ao(s) trecho(s) da(s) SD(s) por índices numéricos. Em (7), a S1 do sumário (C2), alinhada à S1 do D2, tem 2 casos de generalização delimitados e indexados.

(7) Sumário

Em uma operação chamada "Operação Dominó", a Polícia Federal prendeu na manhã desta sexta-feira 23 [pessoas suspeitas]¹ de envolvimento em esquema da Assembleia Legislativa do Estado de Rondônia para [desvio de recursos públicos e influência indevida]² sobre Poder Judiciário, Ministério Público, Tribunal de Contas e Poder Executivo do Estado. (S1)

Texto-fonte

A Polícia Federal desencadeou na manhã desta sexta-feira em Rondônia, a "Operação Dominó", destinada a prender uma [quadrilha de altos funcionários públicos]¹ especializada em [desvio de recursos públicos, corrupção, peculato, extorsão, lavagem de dinheiro e venda de sentenças judiciais.]² (S1_D2)

Assim, registra-se explicitamente que o alinhamento da SSum à SD foi rotulado como “generalização” porque a SSum possui 2 casos de generalizados. Em um desses casos, o trecho [pessoas suspeitas] foi interpretado como generalização de [quadrilha de altos funcionários públicos] e, por isso, ambos foram anotados com o mesmo índice (“1”).

Quanto à natureza dos trechos, ressalta-se que a maioria deles consiste em unidades lexicais (p.ex.: [Peterka] em (8)) ou sintagmas (p.ex.: [um perito aposentado] em (8)). Em alguns casos, as sentenças inteiras foram delimitadas, posto que a SSum veiculava informação inferida do conteúdo global das SDs, como ilustrado em (9).

(8) **Sumário**

Por outro lado, de acordo com [um perito aposentado]1, o fato de Congonhas ter uma pista menor não significa uma causa direta para o acidente. (S10)

Texto-fonte

De acordo com [Peterka]1, o fato de Congonhas ter uma pista menor não significa uma causa direta para o acidente. (S4)

(9) **Sumário**

[No primeiro tempo houve outras jogadas, mas sem gols, ainda que o Brasil tenha se esforçado]1. (S5)

Textos-fonte

[Após um zagueiro equatoriano se atrapalhar, o ligeiro centroavante Vágner Love roubou a bola e chutou na trave, aos 15min.]1 (S10_D1)

[Robinho, Ronaldinho e Kaká se esforçaram, mas não criaram muitas jogadas.]1 (S15_D1)

[A primeira jogada só saiu aos nove minutos, após boa triangulação entre Kaká, Vágner Love e Robinho, que chutou par fora.]1 (S6_D4)

[Aos 14, Juan lançou do campo de defesa, a zaga falhou e Love acertou a trave.]1 (S7_D4)

[Aos 30, Robinho arriscou de longe e obrigou o goleiro Viteri a fazer boa defesa.]1 (S11_D4)

[A última chance de gol no primeiro tempo veio dos pés de Ronaldinho, em cobrança de falta que passou muito perto da trave.]1 (S12_D4)

Ao total, identificaram-se 53 casos de generalização. Especificamente, das 40 SSum alinhadas por generalização, 11 delas possui apenas 2 caso, 1 delas possui 3 casos e as demais 28 possuem apenas 1 caso de condensação de conteúdo por generalização.

4.2. Identificação, anotação e descrição das operações de “recorta e cola”

Para essa tarefa, partiu-se da tipologia de operações identificadas na literatura, a saber: redução sentencial, combinação sentencial, transformação sintática e paráfrase lexical (cf. 2.3). Dessas, apenas a paráfrase lexical não ocorreu no *corpus*.

Ademais, outras operações distintas de reescrita foram identificadas e nomeadas: lexicalização, substituição lexical, substituição sintagmática, substituição causa-efeito,

substituição numérica e substituição dêitica. Para cada uma delas, propõe-se uma etiqueta que foi utilizada para anotar os trechos das SSum e das SDs previamente delimitados e indexados. Tais etiquetas foram compostas pelas letras iniciais em maiúsculo de cada palavra que constitui a denominação da operação. Para “redução sentencial”, por exemplo, a etiqueta especificada foi **RS**.

O formato de anotação preestabelecido foi o seguinte: todas as operações de “recorta e cola” referentes ao caso de generalização delimitado e indexado foram agrupadas ao fim, delimitadas entre si pela barra de divisão (/). Assim, o formato genérico da anotação para *n* operações foi dado por [trecho]índice_operação1/operação2/.../operação_n. Em (10), ilustra-se essa anotação.

(10) **Sumário**

Por outro lado, de acordo com [um perito aposentado]1_SS/CS, o fato de Congonhas ter uma pista menor não significa uma causa direta para o acidente. (S10)

Texto-fonte

De acordo com [Peterka]1_SS/CS, o fato de Congonhas ter uma pista menor não significa uma causa direta para o acidente. (S4)

Uma vez anotada, cada operação foi descrita com o objetivo de registrar a interpretação que resultou na sua identificação e anotação. Por exemplo, a operação “substituição sintagmática” (SS) identificada em (10) foi descrita ou interpretada como “substituição de nome próprio de pessoa pela função/cargo”. Sobre a operação “combinação sentencial” (CS), também em (10), o comentário registrado foi: “a expressão *perito aposentado* advém de outro texto-fonte da mesma coleção”.

Ao final da anotação, 9 operações distintas foram encontradas no *corpus*. A seguir, cada uma delas é definida e exemplificada com ocorrências do *corpus*.

Operação	Transformação sintática	
Etiqueta	TS	
Definição	Operação de “recorta e cola” pela qual uma SSum (ou um trecho dela) é formulada com base na transformação sintática da(s) SD(s) (ou de parte dela); essa transformação consiste na mudança de modalização (exemplo 1) ou de sujeito (exemplo 2).	
Exemplo1	Sumário	[Quatro pessoas teriam ficado feridas.]1
	Texto-fonte	No começo da rebelião [quatro pessoas ficaram feridas, entre elas uma auxiliar de enfermagem e um agente de polícia que trabalham no presídio.]1
Exemplo2	Sumário	Para o segundo turno, as [intenções de voto para Lula caíram]1 quando se compara com os candidatos Geraldo e Heloisa.
	Tex-fonte	De acordo com a CNI/Ibope, num possível segundo turno, [Lula perderia 10 pontos de vantagem]1 para Alckmin.

Operação	Redução sentencial	
Etiqueta	RS	
Definição	Operação de “recorta e cola” pela qual uma SSum (ou um trecho dela) é escrita a partir da remoção ou exclusão de palavras, expressões, sintagmas e orações de uma sentença-fonte.	
Exemplo	Sumário	[Metade dos vôos] ¹ programados entre 6 e 8 horas foram cancelados.
	Texto-fonte	[Quase metade dos vôos] ¹ previstos para decolar na manhã desta terça-feira (24) no Aeroporto de Congonhas, na Zona Sul de São Paulo, foi cancelada, de acordo com informações da Infraero.

Operação	Lexicalização	
Etiqueta	L	
Definição	Operação de sumarização pela qual um conceito específico, expresso originalmente em uma SD por uma expressão multipalavra, é lexicalizado no sumário, sendo expresso por unidade lexical de significado mais genérico.	
Exemplo	Sumário	[Washington] ¹ e a Comissão Européia também pedem uma solução diplomática.
	Texto-fonte	[O presidente dos EUA, George W. Bush] ¹ , pediu nesta quarta-feira que o Exército turco busque uma solução diplomática para a questão.
Comentário	O texto do sumário não contém material linguístico do trecho da sentença do texto-fonte ao qual foi alinhado, mas sim de outra sentença da mesma coleção.	

Operação	Combinação sentencial	
Etiqueta	CS	
Definição	Operação de “recorta e cola” pela qual a sentença de um sumário multidocumento é formulada pela combinação de material linguístico de duas ou mais SDs.	
Exemplo	Sumário	[A seleção se mostrou contrariada pelas vaias] ¹ , e [alguns jogadores] ² [manifestaram sentir falta do jogador Ricardinho.] ³
	Texto-fonte	Para essa partida, [os jogadores esperam uma posição diferente da torcida.] ¹ [Novo capitão da seleção, o ponta Giba] ² [deixou escapar, após o jogo, que sente muita falta do levantador Ricardinho] ³ . [Fiquei surpreso porque ninguém espera isso quando joga em casa.] ¹

Operação	Substituição lexical	
Etiqueta	SL	
Definição	Operação de sumarização pela qual um conceito lexicalizado de uma SD é generalização por outra unidade lexical no sumário.	
Exemplo	Sumário	Um [acidente] ¹ envolvendo dois trens, ao norte do Cairo, deixou por volta de 80 mortos e 165 feridos, segundo [fontes policiais e médicas] ² .
	Texto-fonte	Cairo - [O ministro da Saúde egípcio, Hatem El-Gabaly] ² , informou nesta segunda-feira que 57 pessoas morreram e 128 ficaram feridas no [choque] ¹ entre dois trens de passageiros no delta do Nilo, ao norte do Cairo.
Comentário	O texto do sumário não contém material linguístico do trecho da sentença do texto-fonte ao qual foi alinhado, mas sim de outra sentença do texto-fonte.	

Operação	Substituição sintagmática	
Etiqueta	SS	
Definição	Operação de “recorta e cola” pela qual uma unidade lexical com significado específico expressa um uma SD é generalizada por uma unidade sintagmática.	
Exemplo	Sumário	Por outro lado, de acordo com [um perito aposentado] ¹ , o fato de Congonhas ter uma pista menor não significa uma causa direta para o acidente.
	Texto-fonte	De acordo com [Peterka] ¹ , o fato de Congonhas ter uma pista menor não significa uma causa direta para o acidente.
Comentário	A expressão sintagmática genérica é comumente proveniente de um dos textos-fonte da respectiva coleção do sumário.	

Operação	Substituição causa-efeito	
Etiqueta	SCE	
Definição	Operação de sumarização por meio da qual a causa de um evento é generalizada pelo seu efeito.	
Exemplo	Sumário	Os dois maiores rios do Reino Unido, Severn e Tâmsa, ameaçam transbordar nesta segunda, [agravando ainda mais a situação] ¹ .
	Texto-fonte	Cerca de 350 mil britânicos do sudoeste do país podem [ficar sem abastecimento de água] ¹ devido à ameaça de mais transbordamento dos rios Severn e Tâmsa.

Operação	Substituição dêitica	
Etiqueta	SD	
Definição	Operação de sumarização em que a expressão generalizada “aponta” para o contexto situacional, ou seja, seu referente está na situação; a exata localização (temporal) do fato requer que os interlocutores conheçam o referente que está fora do contexto linguístico (cotexto).	
Exemplo	Sumário	[Está se encerrando] ¹ o prazo para que os deputados acusados de participar do esquema dos sanguessugas renunciem para escapar da abertura de processo por quebra de decoro parlamentar que será instaurado pelo presidente do Conselho de Ética.
	Texto-fonte	Os deputados acusados de envolvimento na máfia dos sanguessugas têm [até a meia-noite desta segunda-feira] ¹ para renunciar aos mandatos se quiserem escapar da cassação. [Termina hoje, às 20 horas] ¹ , o prazo para que os deputados acusados de participar do esquema dos sanguessugas renunciem para escapar da abertura de processo por quebra de decoro parlamentar.

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

Tabela 2: Quantificação numérica das operações de “recorta e cola” nos alinhamentos com generalização.

Operação	Quantidade absoluta	Porcentagem
Substituição sintagmática	17	24,3
Substituição lexical	16	22,8
Redução sentencial	12	17,2
Combinação sentencial	11	15,7
Transformação sintática	6	8,5
Lexicalização	3	4,3
Substituição numérica	3	4,3
Substituição causa-efeito	1	1,45
Substituição dêitica	1	1,45
TOTAL	70	100%

Quanto aos dados quantitativos da Tabela 2, cabem aqui algumas observações:

- A operação “substituição sintagmática” é a mais frequente nos casos de generalização do CSTNews, com 24,5% das ocorrências, seguida de perto pela “substituição lexical”, com 22,8% de frequência;
- Os índices de ocorrência das operações de “redução sentencial” e “combinação sentencial” são próximos, ou seja, 17,2% e 15,7%, respectivamente;
- A operação de “transformação sintática” ocupa um lugar intermediário no ranque, com 8,5% dos casos;
- Os operações de “lexicalização” e “substituição numérica” apresentam mesma frequência, 4,3%;
- O mesmo ocorre com “substituição numérica” e “substituição causa-efeito”, ambas com frequência de 1,45%;
- Dos 11 casos de “combinação sentencial”, 6 deles se caracterizam pelo fato de a SSum estar alinhada a apenas uma SD; nesses caso, o trecho generalizado do sumário apresenta material proveniente de outra S do mesmo texto-fonte da sentença alinhada ou outro texto-fonte da mesma coleção;
- Em 5, dos 11 casos de “combinação sentencial”, os trechos generalizados apresentam material linguístico proveniente dos trechos das duas ou mais sentenças-fonte às quais foram alinhados;
- Os trechos generalizados por meio da operação de “substituição sintagmática” ora são expressões provenientes dos próprios texto-fonte ora são expressões novas (ou seja, que não ocorreram em nenhum dos textos-fonte da coleção);
- Os trechos generalizados por meio das operações de “substituição lexical” ora são unidades provenientes dos próprios texto-fonte ora são unidades novas (ou seja, que não ocorreram em nenhum dos textos-fonte da coleção);
- Nos 3 casos de “lexicalização”, as unidades lexicais genéricas utilizadas nos sumários são provenientes dos textos-fonte, tendo ocorrido em outras sentenças dos textos que não as alinhadas.

No Quadro 4, na sequência, apresenta-se essa distribuição em função dos *clusters* e das SSum alinhadas.

Caso	Cluster	SSum	Sentença_Documento					Transformação sintática	Redução sentencial	Lexicalização	Combinação sentencial	Substituição				
			SD1	SD2	SD3	SD4	SD5					Lexical	Sintagmática	Causa-efeito	Numérica	Dêitica
1	C2	6	8					2		1	1					
2	C3	10		4							1		1			
3	C9	1		1					1		1		1			
4	C9	2	2,4	2,3	3						1		1			
5	C9	3	8								1		1			
6	C9	5	12	15	16						1		1			
7	C13	5			12							1				
8	C14	1	1									1	1			
9	C16	1	1	1												1
10	C18	2			2						1	1				
11	C18	6		16								1				
12	C19	4	5										1			
13	C22	2		2			1,2		1							
14	C23	2	3										1			
15	C23	3	11											1		
16	C25	7				14			1			1				
17	C26	1	27										1			
18	C26	9	25										1			
19	C27	2	25, 27, 28, 29	12,15		1,13,19,24,26,27					1					
20	C27	5	10,15			6,7,11,12					1					
21	C29	1	1					1	1							
22	C29	2		2	2				1			1				
23	C32	4				17						1				
24	C34	3			2						1				1	
25	C34	7		14									1			
26	C35	1					2	1							1	
27	C35	4	10									1				
28	C35	7					5		1	1			1			
29	C37	3		4				1							1	
30	C37	5		8				1	1							
31	C39	3				5							1			
32	C40	2				1			1			1				
33	C42	2		5								1				
34	C42	4		4,2					1			1				
35	C46	1		7								1				
36	C47	6			5					1						
37	C48	6		16								1				
38	C48	7	17,2	17					2		1	2	2			
39	C50	1				1			1				1			
40	C50	5				2,4						1	1			
			TOTAL					6	12	3	11	16	17	1	3	1

Quadro 4: Tabulação das operações de “recorta e cola” identificadas nos casos de generalização do CSTNews.

4.3. Organização e registro dos dados de análise

Todo o processo de identificação das operações de “recorta e cola” referentes à generalização no CSTNews foi feito por meio do *Microsoft Excel*. Para tanto, criou-se uma tabela no formato *Excel* como a ilustrada no Quadro 5.

Nela, observa-se que, na coluna A, registrou-se o número do caso em que uma SSum foi alinhada a uma ou mais SDs, que vai de 1 a 40. Na coluna B, tem-se a SSum e, nas colunas C e D, as sentenças dos textos-fonte 1 e 2 às quais SSum foi alinhada. Na coluna F, foram registradas as interpretações ou os comentários a respeito de cada operação de “recorte e cola”. Ressalta-se que os trechos sentenciais diretamente envolvidos nas generalizações estão delimitados e indexados, assim como estão anotados pelas etiquetadas que indicam as operações de “recorta e cola”. Os comentários referentes a cada operação também estão indexados aos trechos sentenciais.

A	B	C	D	E	F
Caso	Cluster	SSum	D1	D2	Interpretação/Comentário
1	C2	De acordo com a CNI/Ibope, num possível segundo turno, [Lula perderia 10 pontos de vantagem] 1_TS/L/CS para Alckmin. (S8)	Para o segundo turno, as [intenções de voto para Lula caíram] 1_TS/L/CS quando se compara com os candidatos Geraldo e Heloisa. (S6)	X	1_TS: Mudança de modalização e de sujeito. 1_L: Inferência “perder 10 pontos de vantagem” é “cair”. 1_CS: Inserção da expressão “intenção de voto” proveniente de outro texto-fonte da coleção.
2	C3	Em uma operação chamada "Operação Dominó", a Polícia Federal prendeu na manhã desta sexta-feira 23 [pessoas suspeitas] 1_SS/CS de envolvimento em esquema da Assembleia Legislativa do Estado de Rondônia para [desvio de recursos públicos e influência indevida] 2_RS/SS sobre Poder Judiciário, Ministério Público, Tribunal de Contas e Poder Executivo do Estado. (S1)	X	A Polícia Federal desencadeou na manhã desta sexta-feira em Rondônia, a "Operação Dominó", destinada a prender uma [quadrilha de altos funcionários públicos] 1_SS/CS especializada em [desvio de recursos públicos, corrupção, peculato, extorsão, lavagem de dinheiro e venda de sentenças judiciais.] 2_RS (S1)	1_SS/CS: Emprego da expressão “pessoas suspeitas”, que é proveniente de outra sentença do D2. 2_RS: Emprego de parte da expressão do texto-fonte (“desvio de recursos públicos”). 2_SS: Emprego da expressão “influência indevida”, que generaliza “corrupção, peculato, extorsão, lavagem de dinheiro e venda de sentenças judiciais.”

Quadro 5: Organização e registro dos dados de análise.

4.4. Análise das operações de “recorta e cola” envolvidas na generalização

A partir dos dados do Quadro 5, analisaram-se todas as ocorrências de cada uma das operações de “recorta e cola” identificadas no *corpus* com o objetivo de se identificar padrões ou regularidade nos processos de generalização. Tais regularidades subsidiariam a proposição das estratégias descritas na próxima subseção. A seguir, descreve-se a análise feita para cada uma das operações. A descrição dessas análises segue a ordem inversa da apresentada na Tabela 2. Em outras palavras, descrevem-se as operações menos frequentes e depois as mais frequentes do *corpus*.

4.4.1. Generalização por substituição dêitica

Na única ocorrência de generalização no CSTNews por meio de uma “substituição dêitica”, os trechos dos textos-fonte “(x tem) até a meia-noite desta segunda-feira (para y)” e “(o prazo) termina hoje, às 20 horas” caracterizam-se, respectivamente, por: (i) ocorrência da preposição “até” seguida de período do dia (“meia-noite”) e dia da semana (“segunda-feira”), e (ii) ocorrência do verbo “terminar” seguido de dia da semana (“hoje”) e hora (“às 20 horas”). O trecho do sumário resultante do processo de generalização, “(o prazo) está se encerrando”, caracteriza-se pela ocorrência de uma expressão formada pelo verbo de ligação “estar” e pelo verbo no gerúndio “encerrar(-se)”. Percebe-se que o uso do gerúndio indica que “o prazo” ainda não chegou ao fim. Em outras palavras, o término do prazo nos trechos dos textos-fonte é explicitamente expresso pelas informações de dia e hora, enquanto que, no sumário, essa informação foi transposta por meio de um gerúndio, que indica um processo (ainda não terminado).

4.4.2. Generalização por Substituição causa-efeito

Na única ocorrência de generalização por meio de uma “substituição causa-efeito”, o trecho do texto-fonte “ficar sem abastecimento de água” foi generalizado para “agravando ainda mais a situação”. Nesse caso, trata-se de um *cluster* sobre um transbordamento de dois rios importantes do Reino Unido que gerou muitos problemas para determinada região britânica, sendo que o fato de os cidadãos ficarem sem água “agrava a situação”. Novamente, assim como acontece com a generalização por “substituição dêitica”, trata-se de um único caso no *corpus* e, por isso, a estratégia/regra proposta é pontual, ou seja, aplicável apenas para esse caso.

4.4.3. Generalização por Substituição numérica

Os 3 casos de generalização por “substituição numérica” são bastante distintos, sendo que 2 delas caracterizam-se por inúmeras transformações do material linguístico. Para apenas um dos 3 casos, foi possível especificar/delimitar a operação de substituição numérica. Trata-se da generalização de “16 crianças e 14 adultos” para “30 refêns”. Observa-se que o número de adultos e crianças envolvidos em um incidente foi somado ($16+14=30$) e a informação sobre as pessoas foi generalizada para “refêns”.

4.4.4. Generalização por Lexicalização

Dos 3 casos de generalização por “lexicalização”, foi possível especificar/delimitar a operação de lexicalização para um dos casos. O caso diz respeito à generalização de “O presidente dos EUA, George W. Bush” para “Washington”. Aqui, observou-se que a expressão “o presidente de x (país), nome próprio do presidente” pode ser generalizada pelo nome da cidade-sede do governo que o presidente representa.

4.4.5. Generalização por Transformação sintática

Dos 6 casos de “transformação sintática”, 5 deles se caracterizam pela mudança de sujeito e 1 deles consiste na mudança de modalização. Mesmo diante de um número maior de casos em relação às demais operações até então analisadas, não foi possível observar regularidades nos casos. Assim, no Quadro 7, tem-se 3 estratégias para “mudança de sujeito” e 1 estratégia para “mudança de modalização”.

4.4.6. Generalização por Combinação sentencial

A “combinação sentencial” é uma das operações de generalização que se mostrou mais complexa, posto que a origem da expressão genérica é múltipla, ou seja, o trecho que expressa a ideia generalizada resulta do conteúdo veiculado por várias sentenças-fonte. Dos 11 casos identificados, foi possível especificar/delimitar a operação de combinação sentencial em 3. Dois dos casos dizem respeito a textos sobre política. Observou-se que listas de nomes de políticos ou de cargos públicos veiculadas em mais de uma sentença-fonte foram generalizadas para “alto escalão político”, que se originou de outras sentenças dos textos-fonte que continham as expressões “altos funcionários públicos” e “alto escalão do governo”. Além disso, listas de nomes de estados brasileiros veiculados por mais de uma sentença-fonte foram generalizadas para “vários estados”. Nos textos de esporte, observou-se que diferentes jogadas bonitas de futebol (p.ex.: “chutar no ângulo”, “fazer um golazo”, etc.), realizadas por jogadores específicos (x e y), sendo essas jogadas veiculadas em mais de uma sentença-fonte, foram generalizadas para “*belas atuações de x e y* ”.

4.4.7. Generalização por Redução sentencial

Para 3 dos 12 casos de “redução sentencial”, foi possível especificar/delimitar a operação de redução sentencial. Em um dos casos, houve a generalização da expressão “quase metade de x ” (p.ex.: “quase metade dos vôos”) por “metade de x ”. Para tanto, houve a supressão do “quase” (p.ex.: “[~~quase~~] metade dos vôos”). Em outro caso, a expressão formada por um cargo político e um nome do político foi generalizada pela supressão do cargo político e pela expressão apenas do nome próprio, como de “O presidente do Senado, Renan Calheiros” para “[~~O presidente do Senado~~] Renan Calheiros”. O terceiro caso cuja operação de redução sentencial pode ser delimitada caracteriza-se pela generalização da expressão formada por uma função profissional e um nome próprio pela (i) supressão da função e (ii) expressão apenas do nome próprio, como de “advogado das supostas vítimas, Ray Boucher” para “advogado [das supostas vítimas]”.

4.4.8. Generalização por Substituição lexical

Dos 16 casos de generalização por “substituição lexical”, foi possível especificar/delimitar a operação de reescrita em 5 deles. Na primeira delas diz respeito à substituição do nome de cidade pelo nome do estado do qual ela faz parte (p.ex.: de “[comprar uma emissora de rádio e TV, em] Maceió” por [duas emissoras de rádio de] Alagoas). Na segunda, a generalização se dá pela troca de uma expressão numérica de quantidade por um pronome indefinido (p.ex.: de “[Cerca de] 300 edifícios” para “diversos [edifícios]”). Na terceira, observou-se a generalização por meio da substituição do nome próprio de pessoa pela função que ela desempenha (p.ex.: de “[os jogadores mal encaravam] Bernardinho para “[os jogadores mal encaravam o] treinador”). A quarta diz respeito à generalização um tipo de acidente pelo seu hiperônimo (p.ex.: de “choque entre dois trens de passageiros [no delta do Nilo, ao norte do Cairo]” para “[um] acidente [que envolveu]”). E, por fim, na quinta, a generalização por “substituição lexical” se caracteriza pela troca de uma expressão que indica parte por uma expressão que indica o todo (p.ex.: de “[todo] salão da faculdade” para “edifício”).

4.4.9. Generalização por Substituição sintagmática

Dos 17 casos de generalização por “substituição sintagmática”, especificou-se a operação de reescrita em 6 casos. Em dos casos, a operação diz respeito à substituição de um nome próprio por um sintagma preposicional (SPrep) do tipo *pronome possessivo + nome* (=função), como em “[disse] Cahe, [em declarações ao jornal "La Nación"] para “[segundo] seu médico”. Esse caso, aliás, pode ser visto como a transformação da construção “*disse + nome próprio*” para “*segundo SPrep [prep+ nome (= função)]*”, sendo que na primeira tem-se o nome próprio da pessoa e, na segunda, expressa-se a relação (por meio do pronome possessivo “seu”) de quem disse com o personagem central da notícia (no caso, “Maradona”). A segunda se refere à generalização da informação específica sobre a quantidade de chuva que caiu inesperadamente em determinada região. No caso, a quantidade de chuva, expressa em centímetros, foi generalizada pela expressão “(chover) muito acima do esperado”. A terceira se refere ao caso em que a estocagem de uma série de itens (alimento, água, lanterna e vela) necessários para suportar a passagem de um furacão foi substituída pela expressão no gerúndio “(outros estão) preparando-se (para a passagem do furacão)”. Nesse caso, entende-se que “estocar alimento, água, lanterna e vela” é um tipo de preparação para a passagem do furacão. A quarta estratégia é a substituição da expressão “dois terços das autuações” por “irregularidades mais comuns”. Nesse caso, vê-se que a expressão “dois terços” foi generalizada para “mais comuns” e que o termo “autuações” foi generalizado por “irregularidades”. A substituição de “autuações” por “irregularidades” pode ter sido motivada pela relação de causa-efeito, já que as “irregularidades (na declaração de imposto)” causam as “autuações” (efeito). A quinta se refere à generalização de um nome próprio (de pessoa) por um sintagma nominal (SN) que veicula a função/cargo que desempenha (p.ex.: “de [Peterka] para [um perito aposentado]”). Por fim, no sexto

caso cuja operação de substituição sintagmática pode ser sistematizada, observou-se a substituição de uma lista de tipos de crimes administrativos pela expressão mais genérica “crimes administrativos” (p.ex.: de “[pagamentos de serviços, compras e obras supostamente superfaturadas] para [crimes administrativos]”).

5. Proposição de estratégias automáticas de “recorta e cola”

Uma vez delimitadas, as operações de reescrita revelam-se como estratégias humanas de generalização, as quais foram traduzidas em regras no formato lógico *se, então*, tratável por máquina. Especificamente, essas regras codificam as condições (*se*) de ocorrência da generalização 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 em um dos textos-fonte, há um processo de generalização que pode ser aplicado.

No Quadro 6, 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 específica na condição e/ou na ação. Por exemplo, na regra para “substituição dêitica”, a expressão da condição (**se**) deve iniciar pela preposição específica *até*. O restante da condição indica, por exemplo, que a preposição *até* pode ser seguida por qualquer período do dia (p.ex.: meia-noite, meio-dia, etc.) e dia-da-semana (p.ex.: segunda, terça).

Operação “recorta e cola”	Estratégia/Regra	Qt.
Substituição dêitica	Se [<i>até</i> + período do dia + dia] ou [<i>terminar</i> + dia + hora] Então <u>generalizar para</u> [<i>está se encerrando</i>] (* verbo <i>estar</i> + partícula <i>se</i> + verbo <i>encerrar</i> no gerúndio)	1
Substituição causa-efeito	Se [<i>ficar sem abastecimento de água</i>] Então <u>generalizar para</u> [<i>agravando ainda mais a situação</i>]	2
Substituição numérica	Se [número1 + <i>adultos</i> + <i>e</i> + número2 + <i>crianças</i>] Então <u>somar</u> número1 e número2 Então valor total quantificar <i>reféns</i>	3
Lexicalização	Se [<i>presidente</i> de x (=nome de país) + nome próprio] Então <u>generalizar para</u> [cidade-sede do governo]	4
Transformação sintática	Se [x (=nome próprio) <i>perder</i> y (=dado numérico) <i>pontos</i>] Então [<i>intenção de voto para x cair</i>]	5
	Se [<i>a prisão de x</i> (=nome próprio) <i>faz parte da operação</i> y (=nome próprio) Então <u>generalizar para</u> [<i>operação y prendeu w</i> (=dado numérico) <i>suspeitos</i>]	6
	Se [x <i>for libertado</i> {por y}] (* o sujeito da passiva pode não ocorrer na sentença original, mas em outra do mesmo texto-fonte) Então <u>generalizar para</u> [y <i>liberar</i> x]	7

	Se [<i>x ter ficado y</i> (=adjetivo)] Então <u>generalizar para</u> [<i>x ficar y</i> (=adjetivo)]	8
Combinação sentencial	Se [lista de nomes próprios de políticos] ou [lista de cargos públicos] Então <u>generalizar para</u> [<i>alto escalão político</i>]	9
	Se [lista de nomes de estados brasileiros] Então <u>generalizar para</u> [<i>vários estados</i>]	10
	Se [lista de jogadas de futebol bonitas realizadas por <i>x</i> (=nome próprio de jogador)] Então <u>generalizar para</u> [<i>bela atuação de x</i>]	11
Redução sentencial	Se [cargo político + nome próprio de um político] Então <u>generalizar para</u> [nome próprio do político]	12
	Se [profissão + nome próprio do profissional] Então <u>generalizar para</u> [profissão]	13
Substituição lexical	Se [<i>em</i> nome de cidade] Então <u>generalizar para</u> [<i>de</i> nome do estado]	14
	Se [dado numérico (>2) <i>x</i> (= nome contável)] Então <u>generalizar para</u> [<i>diversos x</i> (= nome contável)]	15
	Se [nome próprio de jogador/treinador] Então <u>generalizar para</u> [cargo ou posição que desempenha no time]	16
	Se [um tipo de acidente] (* p.ex.: <i>choque de trens</i>) Então <u>generalizar para</u> [<i>acidente</i>] (hiperônimo)	17
	Se [parte de um edifício] (* p.ex.: <i>salão da faculdade</i>) Então <u>generalizar para</u> [<i>edifício</i>] (* que representa o todo da parte)	18
Substituição sintagmática	Se [<i>disse</i> + nome próprio] Então <u>generalizar para</u> [<i>segundo seu</i> + nome (=profissão)]	19
	Se [<i>x</i> (=pessoas) <i>estocar</i> (alimentos, água)] Então <u>generalizar para</u> [<i>x</i> (=pessoas) <i>preparam-se para</i>]	20
	Se [<i>dois terços</i> (de <i>x</i>)] Então <u>generalizar para</u> [(<i>x</i> ou hiperônimo) <i>mais comuns</i>]	21
	Se [<i>autuações</i> (de contribuintes)] Então <u>generalizar para</u> [<i>irregularidades</i>]	22
	Se [nome próprio] Então <u>generalizar para</u> [SN (artigo + nome (= profissão))] (* p.ex.: <i>um perito</i>)	23
	Se [lista de crimes administrativos] (* p.ex.: <i>pagamentos superfaturados</i>) Então <u>generalizar para</u> [<i>crimes administrativos</i>] (hiperônimo)	24

Quadro 6: Formalização das estratégias de generalização em regras no formato lógico.

6. Considerações finais

Ressalta-se que, neste projeto, as 70 operações de “recorta e cola” identificadas e tipificadas nos 40 casos de generalização do CSTNews foram analisadas, gerando estratégias que podem subsidiar a identificação automática de alguns dos casos de generalização do *corpus*.

As estratégias/regras propostas são bastante pontuais, sendo que algumas delas são até mesmo lexicalizadas, ou seja, regras cuja aplicação depende da ocorrência de uma palavra específica. Além disso, as estratégias/regras são dependentes de domínio e, nesse caso, sua aplicação necessita de repositórios de conhecimento especializado e ferramentas de PLN para o reconhecimento e geração automática de expressões linguísticas que codificam, por exemplo, conceitos como “período do dia” (p.ex.: meio-dia, meia-noite, etc.), “nome de país” (p.ex.: Estados Unidos, EUA, Brasil, etc.), etc. Tais repositórios podem consistir em listas de palavras, taxonomias ou ontologias de domínios e as ferramentas podem englobar, por exemplo, reconhecedores de entidades nomeadas.

Por fim, salienta-se que, por meio deste trabalho, sabe-se mais hoje sobre as operações de reescrita envolvidas no processo de generalização no cenário multidocumento do que se sabia antes do início do projeto.

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

BARBOSA, J.P. **Trabalhando com os gêneros do discurso**: relatar: notícia. São Paulo: FTD, 2001.

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, 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. **Sumarização automática multidocumento: seleção de conteúdo com base no Modelo CST** (*Cross-document Structure Theory*). 2010. 86 f. Dissertação (Mestrado) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2010.

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

DIAS-DA-SILVA, B. C. O estudo linguístico-computacional da linguagem. **Letras de Hoje**, Porto Alegre, v. 41, n. 2, p. 103-138, 2006.

DOLZ, J.; SCHNEUWLY, B. **Gêneros orais e escritos na escola**. Campinas, SP: Mercado de Letras, 2004. 278 p. (Tradução e organização: Roxane Rojo; Glaís Sales Cordeiro).

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

JING, H.; MCKEOWN, 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.

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.

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.; PARDO, T.A.S.; RINO, L.H.M. Sumarização automática multidocumento com mapas de relacionamento. In: STIL STUDENT WORKSHOP ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 2, 2011, Cuiabá/MT. **Proceedings...**Cuiabá, 2011, p. 1-3.

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)