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

Geração de subsídios linguísticos para a detecção automática de aspectos informacionais

Vinícius Felix dos Santos
Ariani Di Felippo

NILC-TR-14-06

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, apresentam-se as primeiras investigações linguísticas sobre os aspectos informacionais baseadas em um *corpus* de sumários multidocumento. Os "aspectos" são unidades básicas de informação identificadas nos sumários multidocumento (isto é, produzidos a partir de uma coleção de textos-fonte que tratam de um mesmo assunto) em função de sua categoria. Por exemplo, os sumários de textos da categoria "acidentes naturais" apresentam um conjunto prototípico de aspectos que podem ser codificados pelos rótulos *what*, *when*, *where*, *why*, *who_affected*, *damages* e *countermeasures*. Aqui, descrevem-se as tarefas de: (i) revisão da anotação manual dos aspectos nos 50 sumários humanos multidocumento do *corpus* CSTNews, (ii) revisão das organizações prototípicas dos aspectos propostas para cada uma das categorias do CSTNews ("esporte", "mundo", "dinheiro", "política", "ciência" e "cotidiano") e (ii) investigação de estratégias linguísticas para a detecção automática dos aspectos informações, as quais são centrais para o desenvolvimento de sumarizadores automáticos baseados em aspectos. 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 (2013/13107-8).

1. Introdução

As características dos sumários humanos multidocumento (ou seja, produzidos a partir de uma coleção de textos que abordam um mesmo assunto) começaram a ser investigadas recentemente devido ao interesse pela produção automática de sumários linguisticamente motivados.

Destacam-se as pesquisas sobre os sumários jornalísticos do tipo informativo e genérico, os quais veiculam, de forma concisa e coerente/coesa, o conteúdo principal de uma coleção de notícias jornalísticas (que discorrem sobre um mesmo fato) de tal forma que uma audiência genérica pode dispensar a leitura das mesmas (MANI, 2001).

Essas pesquisas tem sido desenvolvidas no âmbito da Sumarização Automática Multidocumento (SAM), subárea do Processamento das Línguas Naturais (PLN) em que se busca automatizar a produção de sumários a partir de mais de um texto sobre mesmo assunto, e tem evidenciado que os sumários em questão são comumente compostos por (NENKOVA, 2006; CAMARGO, 2013, RASSI et al., 2013): (i) as informações que se localizam na parte inicial dos textos-fonte; (ii) a informação mais redundante da coleção, pois esta é tida como a mais relevante, (iii) informações provenientes de um dos textos-fonte da coleção em específico, entre outras. Nesse cenário, o trabalho de Camargo (2013) destaca-se por ser o primeiro a investigar diversas propriedades um *corpus* em português. É por causa dele que hoje se sabe mais sobre os sumários do único *corpus* multidocumento de referência do português, o CSTNews (CARDOSO et al., 2011).

Outros trabalhos, também desenvolvidos no âmbito da SAM, têm observado empiricamente que sumários multidocumento jornalísticos expressam conjuntos específicos de "aspectos" (isto é, unidades básicas de informação) em função de sua categoria (p.ex.: AFANTENOS et al., 2004, 2008; ZHOU et al., 2005; OWCZARZAK, DANG, 2011; e LI et al., 2011). Owczarzak e Dang (2011), por exemplo, sugerem que sumários de textos da categoria "acidentes naturais" apresentam os aspectos: *what*, *when*, *where*, *why*, *who_affected*, *damages* e *countermeasures*.

Neste trabalho, 4 tarefas principais foram realizadas: (i) revisão da anotação manual dos aspectos nos 50 sumários humanos multidocumento do *corpus* CSTNews (CARDOSO et al., 2011); (ii) análise da anotação automática dos microaspectos por um reconhecedor de papéis temáticos e (iii) especificação de estratégias linguísticas para a identificação automática de alguns microaspectos. As estratégias de (iii) são centrais para o desenvolvimento de sumarizadores automáticos baseados em aspectos.

Na Seção 2, descrevem-se os principais trabalhos que tratam os aspectos informações. Na Seção 3, descrevem-se as tarefas (i), (ii), (iii) realizadas neste trabalho. Na Seção 4, tecem-se considerações finais sobre as tarefas e os resultados gerados por elas.

2. Revisão da literatura

A revisão da literatura englobou o estudo da origem e definição dos aspectos informacionais, da sua relação com outros construtos teóricos (isto é, "papéis temáticos" e "movimentos retóricos") e da anotação manual dos mesmos no CSTNews, assim como da estrutura da "notícia jornalística".

Com isso, buscou-se especificar estratégias linguísticas que possam fomentar a anotação automática dos aspectos em textos jornalísticos em português.

2.1. A Sumarização Guiada, os aspectos e a estrutura das notícias

O interesse pelos aspectos informacionais em sumários multidocumento produzidos a partir de notícias jornalísticas surgiu recentemente, particularmente na edição de 2010 da TAC[1] (do inglês, *Text Analysis Conference*), que é a principal conferência e competição científica dedicada à Sumarização Automática.

Na TAC de 2010, Owczarzak e Dang (2011) propuseram a tarefa de construção de sumários automáticos multidocumento denominada "Sumarização Guiada" (SG) (do inglês, *Guided Summarization*). Nela, os aspectos são utilizados para "guiar" a construção de sumários em função da categoria (ou domínio) a que os textos-fonte (isto é, notícias jornalísticas) pertencem. O argumento para a proposição dessa tarefa é que, ao se conhecer os aspectos que devem constar nos sumários de cada categoria textual, é possível construir sumários automáticos mais similares aos manuais quanto à informatividade e coerência.

Para tanto, os seguintes aspectos foram delimitados para cada uma das categorias cobertas pela competição:

- Acidentes e desastres naturais: fato, data, lugar, razões do acidente/desastre, entidade afetada, danos e esforços de resgate/contramedidas.
- Ataques: fato, data, lugar, entidade afetada, danos, criminosos e esforços de resgate/contramedidas.
- Saúde e segurança: problema, entidade afetada, como foi afetado, por que o problema ocorre e contramedidas.
- Recursos naturais ameaçados: descrição do recurso, importância do recurso, ameaças e contramedidas.
- Julgamentos e investigações: quem está sob investigação, quem está investigando ou processando, o porquê da investigação, acusações, sentença/consequência, reações dos acusados.

Para representar os aspectos, Owczarzak e Dang (2011) propuseram etiquetas em inglês. Assim, os autores estabeleceram que os sumários da categoria "acidentes e desastres naturais", por exemplo, apresentam os aspectos *what* (fato), *when* (data), *where* (lugar), *why* (razão ou motivo), *who_affected* (entidade afetada), *damages*

[1] http://www.nist.gov/tac

(danos) e *countermeasures* (contramedida), e os sumários da categoria "recursos em extinção ou ameaçados" apresentam os aspectos *what* (fato), *importance* (importância), *threats* e *countermeasures*. Uma vez feita essa identificação, os autores propuseram que os sumários automáticos fossem gerados de modo a contemplar esses aspectos.

Com base da SG e na disponibilização pela TAC de *corpora* em inglês anotados com aspectos, muitos estudos foram desenvolvidos. Steinberger et al. (2010), por exemplo, realizaram análises semânticas para identificar os aspectos relevantes que ocorrem em notícias das categorias "evento violento" e "desastres naturais e artificiais". Li et al. (2011), por sua vez, compilaram aspectos de sumários da Wikipédia. Mesmo antes da TAC 2010, alguns trabalhos já apresentavam abordagens semelhantes: Afantenos et al. (2004, 2008) focaram os aspectos em sumários de esportes e Zhou et al. (2005) estudaram os aspectos em sumários biográficos.

Para o português, o estudo pioneiro dos aspectos em sumários multidocumento com vistas à geração de conhecimento para a SAM utilizou o *corpus* CSTNews (CARDOSO et al., 2011), composto por 50 coleções de textos pertencentes às categorias: "esporte", "mundo", "dinheiro", "política", "ciência" e "cotidiano" (Figura 1). Cada coleção engloba de 2 a 3 textos, cada um deles produzido por uma agência distinta (*Folha de São Paulo*, *Estadão*, *O Globo*, *Gazeta do Povo* e *Jornal do Brasil*), sobre uma mesma notícia. Assim, o *corpus* é composto por coleções das seguintes categorias.

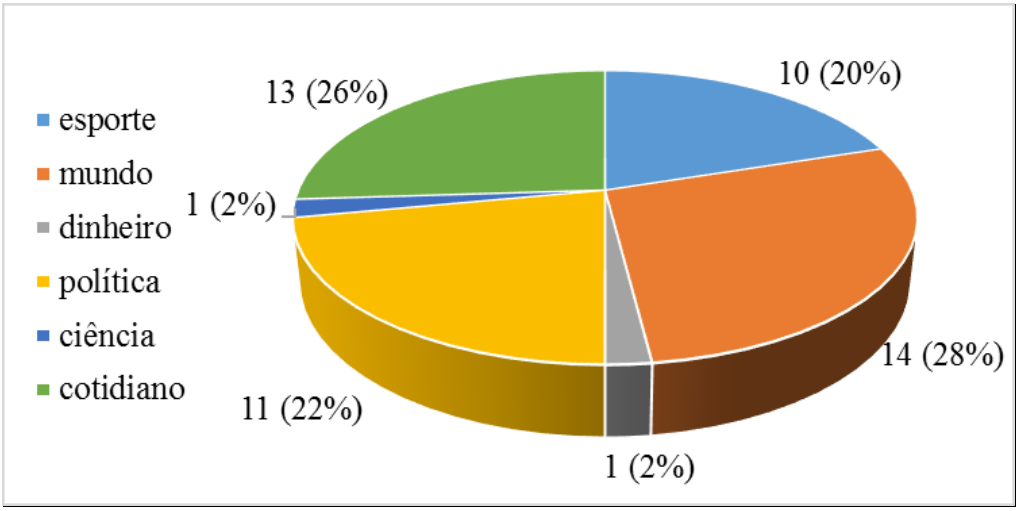

Figura 1: Distribuição das coleções do CSTNews em função das categorias.

Além dos textos-fonte crus (sem nenhum tipo de anotação), cada coleção possui:

a) sumários manuais (*abstracts*) monodocumento, ou seja, um para cada texto-fonte;
b) 1 sumário manual multidocumento, ou seja, produzido por humanos a partir da reescrita do conteúdo dos textos da coleção;

c) 1 sumário automático multidocumento, isto é, produzido por um método particular de SAM;
d) 1 extrato manual multidocumento, ou seja, produzido por humanos a partir de segmentos inalterados dos textos-fonte;
e) versões anotadas, em nível discursivo, de cada um dos textos-fonte com base na *Rhetorical Structure Theory* (RST) (MANN, THOMPSON, 1987) e na *Cross-document Structure Theory* (CST) (RADEV, 2000);
f) versões anotadas, em nível morfossintático e sintático, de cada texto;
g) versão com anotação semântica de cada texto-fonte, especificamente com a anotação dos sentidos/conceitos subjacentes ao substantivos mais frequentes da coleção;
h) versão com anotação de subtópicos para cada texto;
i) alinhamento/associação das sentenças dos sumários manuais às sentenças dos textos-fonte que lhes deram origem, etc.

Assim, o CSTNews é um recurso linguístico-computacional que tem subsidiado as pesquisas em SAM para o português por prover anotações linguísticas de qualidade, produzidas por especialistas nas diversas teorias/modelos que guiaram as anotações.

Por se tratar da primeira anotação sistemática e em larga escala dos aspectos textuais, alguns critérios iniciais foram definidos, a saber (cf. RASSI et al., 2013):

a) unidade de análise: sentença, pois é bem delimitada e veicula uma ideia completa; assim, dado um sumário *x*, cada sentença (S) de *x* é descrita em função dos aspectos que nela ocorrem, sendo que a identificação dos mesmos decorre do conteúdo global do texto, ou seja, da relação de S com as demais sentenças do sumário;
b) conjunto inicial de aspectos/etiquetas (*tags*): 16 aspectos da TAC 2010, representados por etiquetas em inglês, como *what* (objetos genéricos/eventos), *when* (datas), *where* (locativos), *who* (agentes/pacientes), etc.;
c) nível de identificação dos aspectos: os aspectos podem condizer com o conteúdo sentencial (macroaspectos) ou com conteúdo intrassentencial ou sintagmático (microaspectos);
d) formato de anotação: [sentença]lista_de_aspectos; múltiplos aspectos em uma mesma sentença são indicados pela barra [sentença]aspecto1/aspecto2/, etc.
e) redação de definições para os aspectos identificados.

Com base nesses critérios, os 50 sumários humanos das coleções "mundo", "política", "cotidiano", "esporte", "dinheiro" e "ciência" do CSTNews foram manualmente anotados pelos linguistas computacionais que compõem o grupo de sumarização automática do NILC (Núcleo Interinstitucional de Linguística Computacional)[2], sendo que cada categoria ficou destinada a um subgrupo específico

[2] http://www.nilc.icmc.usp.br

(3 ou 4 membros). Atualmente, o grupo é composto por 5 docentes e mais de 20 alunos, os quais são colaboradores e membros do projeto SUSTENTO.

A Figura 2 ilustra a anotação de um sumário da categoria "esporte". No sumário ilustrado, identificou-se, por exemplo, que:

a) a primeira sentença (S1) expressa um "comentário do autor", representado pela *tag* COMMENT;
b) o trecho "o time do Bernardinho" (sujeito) da S2 expressa a "pessoa responsável pela ação", etiquetado com WHO_AGENT;
c) o trecho "derrotou a Rússia [...] (Polônia)" da S2 expressa o aspecto "fato principal descrito no texto", anotado com WHAT;
d) o trecho "por 3 sets a 1, com parciais de 18/25, 25/23, 28/26 e 25/22" da S2 expressa o "resultado numérico de um fato", codificado na etiqueta SCORE;
e) o trecho "em Katowice (Polônia)" da S2 expressa a "localização geográfica do evento", etiquetado com WHERE.

1[A seleção brasileira masculina de vôlei mostrou mais uma vez sua superioridade na modalidade.**]COMMENT**

2[O time de Bernardinho derrotou a Rússia por 3 sets a 1, com parciais de 18/25, 25/23, 28/26 e 25/22, em Katowice (Polônia).**]WHO_AGENT/WHAT/SCORE/WHERE**

3[Com a vitória, a seleção conquistou seu quinto título consecutivo e o sétimo no total na Liga Mundial, aproximando-se da Itália, ainda a maior vencedora de Ligas Mundiais – soma oito conquistas.**]WHO_AGENT/CONSEQUENCE/COMPARISON/HISTORY**

4[O próximo objetivo da seleção é a medalha de ouro nos Jogos Pan-Americanos do Rio.**]GOAL/SITUATION_EXTRA/WHERE_EXTRA**

Figura 2. Exemplo de anotação dos aspectos textuais no CSTNews.

Seguindo-se a diretriz do formato de anotação, os aspectos presentes em uma sentença foram codificados ao final da sentença e na ordem de ocorrência. Por exemplo, na S2, o aspecto WHO_AGENT foi o primeiro a ocorrer e, portanto, o primeiro a ser anotado, seguido por WHAT, SCORE e WHERE, necessariamente nessa ordem.

No Quadro 1, apresentam-se os 17 aspectos do CSTNews. Dentre eles, há aspectos genéricos (p. ex.: WHO, WHAT, etc.) e específicos de cada categoria (p.ex.: SCORE de "esporte").

Ressalta-se também que todas as etiquetas podem ser especificadas como *extra*. Para isso, basta que os segmentos textuais correspondentes não se refiram ao tópico principal do texto (isto é, WHAT).

Na Figura 2, por exemplo, os aspectos "ocasião em que ocorreu o fato" e "localização geográfica" da S4 correspondem a um evento secundário, daí as etiquetas SITUATION_EXTRA e WHERE_EXTRA.

Quadro 1. Aspectos no *corpus* de sumários manuais multidocumento do CSTNews.

Aspecto/Etiqueta	Definição
COMMENT	Um comentário do autor sobre um fato/evento.
COMPARISON	Dados ou estatísticas diferentes comparando duas ou mais entidades.
CONSEQUENCE	Um fato/evento causado por outro fato/evento.
COUNTERMEASURES	Medidas que visam solucionar/antecipar/impedir problemas relacionados a um fato/evento.
DECLARATION	Um discurso ou fala de alguém ou de uma fonte por citação direta ou indireta.
GOAL	Finalidade/razão para um fato/evento que irá acontecer.
HISTORY	Informação de contexto sobre uma história/um passado relacionado ao fato/evento.
PREDICTION	Informação sobre a factibilidade de fatos/eventos futuros.
SITUATION	Uma ocasião em que ocorreu um fato/evento. Envolve uma transação, um campeonato, um compromisso ou outros tipos de situação em uma data ou local inespecíficos.
WHAT	O fato/evento descrito no texto.
HOW	O modo como um fato/evento ocorreu.
SCORE	O resultado numérico de um fato/evento (score, tempo, distância, etc., sobretudo relativo a esportes).
WHEN	A data/período de tempo (estritamente temporal) da ocorrência de um fato/evento.
WHERE	A localização geográfica ou física de um fato/evento.
WHO_AGENT	A entidade (pessoa ou organização) responsável por causar/provocar a ocorrência de um fato/evento.
WHO_AFFECTED	A entidade (pessoa ou organização) que sofre os efeitos de um fato/evento.
WHY	A explicação de o porquê um fato/evento acontece (ou aconteceu).

No Quadro 2, tem-se os aspectos do CSTNews agrupados em macroaspectos e microaspectos. Aqueles marcados com um asterisco (*) no Quadro 2 podem ocorrer como microaspectos ou macroaspectos.

O aspecto SITUATION foi classificado em geral como macroaspecto por ser mais recorrente como tal no *corpus*, assim como o aspecto HOW foi classificado no geral como microaspecto.

Ainda quanto à classificação entre macro e microaspectos, ressalta-se que os macroaspectos são veiculados por sentenças completas e determinados pela relação semântico-discursiva desta com as demais do texto.

Por exemplo, no sumário da Figura 2, tem-se 6 macroaspectos no total. COMMENT está associado à S1 e isso quer dizer que o conteúdo total da referida sentença é um “comentário” do autor sobre o fato descritos nos textos-fonte. O

mesmo pode ser dito para CONSEQUENCE, COMPARISON e COMPANY, que são veiculados pela S3, e GOAL, expresso por S4.

Quanto a SITUATION, ressalta-se que, na categoria "esporte", esse aspecto é de nível micro, sendo expresso, portanto, por um trecho sentencial específico. No caso do texto da Figura 2, SITUATION está associado ao trecho "nos Jogos Pan-Americanos".

Quadro 2. Os macroaspectos e microaspectos do CSTNews.

Macroaspecto	**Microaspecto**
COMMENT	SCORE
COMPARISION	WHEN
CONSEGUENCE	WHERE
COUNTERMEASURES	WHO_AFFECTED
DECLARATION	WHO_AGENT
GOAL	WHY
HISTORY	HOW (*)
PREDICTION	
SITUATION (*)	
WHAT	

Além da identificação dos aspectos, Rassi et al. (2013) propuseram organizações prototípicas para os sumários com base no estudo do *corpus*. Essa organização foi codificada em *templates*, como o ilustrado no Quadro 3.

Quadro 3. Análise dos aspectos nos sumários de "esporte".

Para todos os sumários	
Em comum	WHO_AGENT, WHAT
No 1º parágrafo	WHO_AGENT, WHAT
Ordenação parcial	1. WHO_AGENT<WHAT
Para a maioria dos sumários	
Em comum	WHO_AGENT, WHAT, SCORE, CONSEQUENCE, SITUATION, COMMENT, WHEN, WHERE
No 1º parágrafo	WHO_AGENT, WHAT, SCORE, CONSEQUENCE, SITUATION, COMMENT, WHEN, WHERE
Ordenação parcial	1. WHO_AGENT<WHAT 2. WHO_AGENT, WHAT<SCORE 3. WHO_AGENT, WHAT<CONSEQUENCE 4. WHO_AGENT, WHAT<SITUATION 5. WHO_AGENT, WHAT<WHERE 6. WHO_AGENT, WHAT, SCORE<CONSEQUENCE

Nele, observa-se que, na categoria "esporte", alguns aspectos são mais comuns que outros e que há uma ordenação parcial entre eles. Como isso não ocorre em todos os

casos, as informações sobre os aspectos foram organizadas nos *templates* em função do que foi observado em "todos os aspectos" e na "maioria deles".

Ao se interpretar os dados do *template,* vê-se que todos os sumários da categoria "esporte" apresentam WHO_AGENT e WHAT no primeiro parágrafo, sendo que WHO_AGENT sempre antecede WHAT (ordenação 1). Para a maioria dos sumários, nota-se que WHO_AGENT e WHAT antecedem aspectos específicos, no caso: CONSEQUENCE, SITUATION, SCORE e WHERE. Ademais, quando WHO_AGENT, WHAT e SCORE coocorrem, eles antecedem CONSEQUENCE.

Quanto à ocorrência de WHO_AGENT e WHAT no primeiro parágrafo, ressalta-se que essa característica é justificada pelo fato de que os textos jornalísticos (no caso, notícias) são estruturados como uma "pirâmide invertida", composto por (LAGE, 2002): (a) título; (b) *lead*, que corresponde ao primeiro parágrafo do texto, e (c) corpo do texto, que abrange os demais parágrafos, os quais desenvolvem os elementos informativos referidos no *lead*. O *lead* é a informação principal, expressa com o intuito de instigar o leitor. Logo no início, o leitor tem o essencial da informação: **quem** (sujeito), **o que** (fato/acontecimento), **quando** (tempo), **por quê** (causa/motivo/finalidade), **como** (modo/maneira) e **onde** (lugar).

A pirâmide invertida é ilustrada pela Figura 3, elaborada com base em Lage (2002).

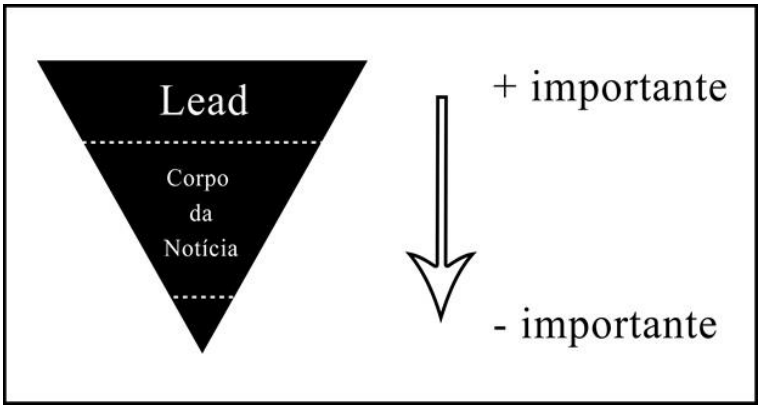

Figura 3: Modelo estrutural da pirâmide invertida.
Fonte: Elaborada com base em Lage (2002)

Com base nas organizações prototípicas ou *templates*, métodos de SAM poderão ser propostos, os quais identificarão nos textos-fonte as sentenças que veiculam as informações previstas nos *templates* para a composição dos sumários automáticos.

Ao final da anotação, obteve-se a distribuição quantitativa dos aspectos no CSTNews como descrita no Quadro 4. Nesse quadro, tem-se essa quantificação sem a distinção entre microaspecto e macroaspecto ou por categoria (esporte, cotidiano, etc.). O número médio de ocorrências de cada aspecto no *corpus* e o desvio padrão[3] também são exibidos na Quadro 4.

[3] O desvio padrão mostra o quanto de variação ou "dispersão" existe em relação à média (ou valor esperado). Um desvio padrão baixo indica que os dados tendem a estar próximos da média; um desvio padrão alto indica que os dados estão espalhados por uma gama de valores.

Quadro 4. Representatividade dos aspectos principais no CSTNews.

Aspecto	Quantidade	Frequência	Média	DesvPad
WHAT_EXTRA	**117**	15%	2.67	2.50
WHO_AGENT_EXTRA	**79**	10%	1.88	2.27
DECLARATION	**59**	7%	1.24	1.46
WHO_AGENT	**55**	7%	1.51	2.91
WHAT	**51**	6%	1.24	1.35
CONSEQUENCE	**43**	5%	1.04	1.80
WHEN	**40**	5%	1.02	1.57
WHO_AFFECTED	**35**	4%	0.82	1.02
WHEN_EXTRA	**34**	4%	0.78	1.22
WHERE	**34**	4%	0.80	0.90
HISTORY	**29**	4%	0.67	1.30
HOW	**25**	3%	0.86	2.87
WHO_AFFECTED_EXTRA	**25**	3%	0.53	1.05
COMMENT	**22**	3%	0.78	2.51
WHERE_EXTRA	**21**	3%	0.51	1.01
COUNTERMEASURES	**18**	2%	0.37	0.80
WHY	**18**	2%	0.41	0.73
PREDICTION	**16**	2%	0.43	0.95
SITUATION	**14**	2%	0.41	0.95
WHY_EXTRA	**14**	2%	0.31	0.58
COMPARISON	**8**	1%	0.20	0.78
SCORE	**7**	1%	0.29	1.03
GOAL	**5**	1%	0.12	0.33
GOAL_EXTRA	**5**	1%	0.10	0.30
SITUATION_EXTRA	**5**	1%	0.20	0.78
COMMENT_EXTRA	**4**	1%	0.16	0.65
CONSEQUENCE_EXTRA	**4**	1%	0.12	0.39
SCORE_EXTRA	**3**	0%	0.12	0.59
COUNTERMEASURES_EXTRA	**1**	0%	0.02	0.14
DECLARATION_EXTRA	**1**	0%	0.02	0.14
HOW_EXTRA	**1**	0%	0.02	0.14
PREDICTION_EXTRA	**1**	0%	0.02	0.14
Totais	**794**	**100%**		

Na sequência, apresenta-se uma breve revisão sobre os papéis temáticos, posto que há similaridade entre estes e alguns microaspectos.

2.2. A relação entre os microaspectos e os papéis temáticos

Como mencionado, uma das diretrizes para a anotação dos sumários do CSTNews foi a de que os aspectos podem condizer com o conteúdo sentencial (macroaspectos) ou com conteúdo intrassentencial ou sintagmático (microaspectos).

Assim definidos, os microaspectos parecem corresponder a certos argumentos projetados pelos predicadores[4] oracionais, os verbos. A todo predicador, está associada uma **estrutura de argumentos** ou **valência**. Os predicadores são itens lexicais semanticamente incompletos que, por isso, precisam necessariamente ligar-se a outros elementos (argumentos – As) para adquirir um valor semântico completo (BORBA, 1996; NEVES, 2000). Os verbos são predicadores oracionais porque a estrutura de argumentos projetada por eles equivale a uma sentença completa.

Segundo a Gramática de Valência (BORBA, 1996), a relação estabelecida entre um predicador e seus As ocorre em três níveis: lógico-semântico, sintático ou morfossintático e semântico. No nível lógico-semântico, caracteriza-se o número de As projetado por um predicador. Desse ponto de vista, um verbo pode projetar de 1 a 4 As, podendo ser monovalente, bivalente, trivalente ou tretavalente. No nível sintático, estabelecem-se as categorias sintagmáticas dos As do predicador. A esse fenômeno, dá-se o nome **quadro de subcategorização** (RAPOSO, 1992) ou valência morfossintática (ou sintática) (BORBA, 1996). No nível semântico, as relações semânticas estabelecidas entre um predicador e seus As são representadas por **papéis semânticos** ou **temáticos** (FILLMORE, 1968). Para exemplificar os níveis de valência, considera-se (1).

(1) O homem quebrou o vidro da janela com uma pedra.

Na sentença (1), o predicador é "quebrou" e os seus argumentos são "o homem", "o vidro da janela" e "com uma pedra". No caso, "quebrou" é um predicador de trivalente. No nível sintático, o argumento "o homem" e "o vidro da janela" são sintagmas nominais e "com uma pedra" é um sintagma preposicional. Das relações semânticas que se estabelecem entre o verbo e os argumentos, tem-se os seguintes papéis temáticos: agente ("o homem"), paciente ("o vidro da janela") e instrumento ("com uma pedra").

Existem vários conjuntos de rótulos para os papéis semânticos, pois nem sempre é simples decidir qual o mais adequado para anotar um argumento. Uma dessas propostas é a do PropBank-Br (DURAN, ALUÍSIO, 2012), projeto que se originou do PropBank, repositório de papéis temáticos para verbos do inglês (PALMER et al. 2005).

No Quadro 5, tem-se o elenco de argumentos e papéis temáticos do PropBank-Br. Os argumentos, com exceção de Arg0 e Arg1, recebem a etiqueta Arg**M** por serem vistos como modificadores de seus predicadores.

[4] Por predicador, entende-se todo elemento que atribui uma determinada propriedade a um certo termo ou estabelece uma relação entre termos, ou seja, uma *predicação* (NEVES, 2000).

Quadro 5. Principais argumentos e papéis temáticos do PropBank-Br.

Argumento	Papel semântico	Definição
Arg0	Agente	aquele que realiza a ação; agente, causador ou experienciador
Arg1	Paciente	aquele que é afetado pela ação, ocorrendo uma mudança de estado
ArgM-ADV	Adverbial	elemento sintático que modifica o verbo
ArgM-CAU	Causa	razão da ação
ArgM-PRP	Propósito	motivação da ação expressa pelo verbo
ArgM-EXT	Quantidade	quantidade da mudança da ação
ArgM-MNR	Maneira	modo pelo qual a ação se realizou
ArgM-DIR	Direção	movimento feito sob algum trajeto
ArgM-LOC	Localização	local em que a ação ocorreu
ArgM-TMP	Temporal	tempo em que a ação ocorreu
ArgM-DIS	Discurso	marcador discursivo
ArgM-MOD	Modal	verbos modais
ArgM-NEG	Negação	negação

Utilizando-se dos rótulos propostos pelo PropBank-Br, tem-se a seguinte anotação para a sentença em (1).

(2) [O homem]**Arg0** quebrou [o vidro da janela]**Arg1** [com uma pedra]**ArgM-MNR**.

Se se comparar os rótulos do projeto PropBank-Br com os microaspectos, já é possível identificar algumas correspondências diretas, as quais estão apresentadas no Quadro 6.

Quadro 6. Correspondência entre microaspectos e papéis temáticos.

Microaspecto	Papel temático
WHO_AGENT	Arg0
WHO_AFFECTED	Arg1
WHEN	ArgM-TMP
WHERE	ArgM-LOC
HOW	ArgM-MNR
WHY	ArgM-PRP
SCORE	ArgM-EXT

Dessa forma, a anotação da sentença em (1), com aspectos, é a descrita em (3).

(3) [O homem]**WHO_AGENT** quebrou [o vidro da janela]**WHO_AFFECTED** [com uma pedra]**HOW**

Seguindo-se o formato do CSTNews, a sentença em (1) seria assim anotada:

(4) [O homem quebrou o vidro da janela com uma pedra]**WHO_AGENT**/ **WHO_AFFECTED/HOW**

Como indicado pelo Quadro 2, SITUATION (isto é, "uma ocasião em que ocorreu um fato/evento") foi classificado como macroaspecto porque assim ocorre com mais frequência. No entanto, ele também ocorre como microaspecto em algumas categorias do CSTNews, como em "esporte". Na sentença em (5) (cf. Figura 2), vê-se que o aspecto SITUATION está expresso no trecho "nos Jogos Pan-Americanos".

(5) [O próximo objetivo da seleção é a medalha de ouro nos Jogos Pan-Americanos do Rio.**]GOAL/SITUATION_EXTRA/WHERE_EXTRA**

Dentre os rótulos do repositório PropBank-Br, no entanto, não há um papel temático correspondente ao aspecto SITUATION.

Considerando-se a similaridade entre os microaspectos e os papéis semânticos, a anotação desses aspectos em *corpus* pode ser realizada com o auxílio das estratégias aplicadas ao reconhecimento automático de papéis temáticos.

Para o português, destaca-se o trabalho de Alva-Manchego (2013), no qual se propôs uma ferramenta automática de "anotação de papéis temáticos" (APS) (em inglês, *Semantic Role Labeling,* SRL) para o português do Brasil. Essa ferramenta, denominada "anotador de papéis semânticos", realiza 3 tarefas: (i) identificação do verbo alvo, (ii) identificação dos argumentos e (iii) classificação dos argumentos. O conjunto de rótulos utilizados para classificar os argumentos é o do PropBank-Br.

Partindo-se de textos anotados em nível sintático pelo *parser* PALAVRAS (BICK, 2000), o APS identifica o papel temático de cada argumento projetado pelo verbo com base em um conjunto de 23 propriedades, denominadas "atributos" (do inglês, *features*), tais como tipologia sintagmática, posição do sintagma na sentença, etc.

A configuração sintagmática e a posição dos argumentos, por exemplo, são características importantes para a identificação dos papéis temáticos porque alguns papéis temáticos emergem da relação do verbo com argumentos de configuração sintagmática específica e que ocorrem mais frequência em determinada posição.

Por exemplo, na sentença "*O homem quebrou o vidro da janela com uma pedra*", as informações de que o argumento "com uma pedra" é um sintagma preposicional e de que se localiza após um sintagma nominal ("o vidro da janela_") no interior de um sintagma verbal ("quebrou o vidro da janela com uma pedra") são dicas relevantes para a seleção adequada do papel semântico **maneira** (ArgM-MNR), ou seja, do aspecto **HOW**. Assim, tais atributos, que subsidiam o APS (ALVA-MANCHEGO, 2013), podem subsidiar a identificação automática da maioria dos microaspectos.

Na sequência, revisa-se brevemente o conceito de "movimento retórico" por se assemelhar aos macroaspectos.

2.3. A relação entre os macroaspectos e os movimentos retóricos

Ao contrário dos microaspectos, um macroaspecto é veiculado por uma sentença completa e se define pela relação desta com as demais do texto. Dessa forma, vê-se aqui certa semelhança com os movimentos retóricos de Swales (1990).

A estrutura de textos científico-acadêmicos tem sido alvo de inúmeras pesquisas.

Em um desses trabalhos, Swales e Feak (2009) identificaram que os resumos ou *abstracts* científicos, por exemplo, apresentam certos *movimentos retóricos* (do inglês, *moves*), isto é, segmentos textuais ou blocos discursivos que desempenham funções específicas nos textos (ou estágios comunicativos).

No caso, tratam-se dos seguintes movimentos retóricos: (i) introdução/contexto, (ii) lacuna, (iii) objetivos ou propósitos, (iv) metodologia, (v) resultados, e (vi) discussão/conclusão.

Para a realização de cada movimento, há estratégias retóricas diversas, ou seja, mecanismos linguísticos que o escritor pode escolher para realizar o propósito comunicativo do movimento do texto como um todo. Esses mecanismos de realização do movimento foram denominados por Swales (1990) *passos* (do inglês, *steps*).

A título de exemplificação, o movimento "apresentar a pesquisa" tem como função apresentar a pesquisa por meio da menção aos autores e objetivos. Assim, esse movimento pode ser constituído pela "referência aos autores (passo 1) ou pela "referência ao objetivo da pesquisa" (passo 2).

Os passos dos movimentos retóricos são materializados na superfície linguística por determinados padrões léxico-gramaticais. Por exemplo, o passo 2 do movimento "apresentar a pesquisa" refere-se à "apresentação dos objetivos" e, para tanto, tem-se os seguintes padrões léxico-gramaticais possíveis em inglês: (i) *this paper presents/ describes/ proposes*; (ii) *the purpose/ objective/ aim/goal of this paper/ study/ work* e (iii) *the objective/ aim/ purpose is to*. Tais padrões também são chamados "expressões formulaicas" (em inglês, *formulatic expressions*).

Para a identificação automática dos movimentos retóricos (ou também zonas argumentativas) em *abstracts* em inglês, tem-se sistemas como o AZEA (do inglês, (*Argumentative Zoning for English Abstracts*) (GENOVES JR. et al., 2007) e o MAZEA (do inglês, *Multi-label Argumentative Zoning for English Abstracts*) (DAYRELL et al., 2012), os quais buscam identificar o(s) movimento(s) retórico(s) corretamente expresso(s) nas sentenças de um texto científico com base em um conjunto de atributos sentenciais, que podem apresentar diferentes valores. Em outras palavras, esses sistemas partem do princípio de que o(s) papel(éis) retórico(s) de uma sentença podem ser "lidos" a partir de características superficiais do texto.

Os atributos e valores utilizados nos sistemas AZEA e MAZEA estão sistematizados no Quadro 7. Nesse Quadro, os atributos são representados por

rótulos em inglês conforme a literatura sobre anotadores automáticos de papéis retóricos. Por exemplo, dada uma sentença, o atributo "tamanho da sentença" (medido em número de palavras) pode ter como valores possíveis: *small*, *medium* e *big*.

Para a construção do MAZEA, por exemplo, os atributos do Quadro 7 referentes às sentenças de dois *corpora* foram descritos de forma semiautomática e as zonas argumentativas foram anotadas de forma manual. Um desses *corpora* é formado por 645 *abstracts* científicos da Física e da Engenharia e o outro é formado por 690 *abstracts* das Ciências da Saúde.

Uma vez descritos e anotados, os *corpora* foram submetidos a um sistema de Aprendizado de Máquina que apreendeu padrões estatisticamente relevantes para a identificação das zonas argumentativas. Com base nesses padrões, o sistema obteve precisão de 66% na identificação dos movimentos retóricos.

Quadro 7: Características sentenciais para a identificação automática dos movimentos retóricos.

Atributo (*Feature*)	**Descrição**	**Valores possíveis**
Tamanho	Tamanho da sentença (em nº de palavras)	*Small, Medium, Big*
Localização	Posição da sentença do texto	*Fir, Sec, Third, Med, Penult, Last*
Tempo	Tempo do 1º verbo finito da sentença	*BaseForm, Gerund, Past, PartPart, Pres3, PresNo3, NoVerb*
Voz	Voz do verbo flexionado	*Active, Passive*
Modal	Se o 1º verbo finito da sentença é ou não modal	*Modal, NoModal ou NoVerb*
Histórico	Categoria da sentença anterior	*Background, Gap, Purpose, Method, Result, Conclusion, None*
Expressões formulaicas	Tipo de expressão padrão contida na sentença	19 tipos de expressões ou *none*
Agente	Tipo de agente contido na sentença	14 tipos de agente ou *none*

Considerando-se a similaridade entre os macroaspectos e os movimentos ou papéis retóricos, a anotação dos macroaspectos em *corpus* pode ser realizada com o auxílio das estratégias aplicadas ao reconhecimento automático dos papéis retóricos. No entanto, para isso, é preciso, por exemplo, elencar as "expressões formulaicas" utilizadas na expressão linguística dos macroaspectos.

Enquanto a tarefa de anotação dos textos ainda não foi realizada, seja ela manual ou semiautomática (p.ex.: via anotadores de papéis temáticos), realizou-se uma análise ou revisão da anotação manual dos aspectos nos sumários multidocumento do CSTNews e das organizações prototípicas, como a ilustrada no Quadro 3. A tarefa de revisão está descrita na sequência.

3. Revisão da anotação dos sumários e das organizações prototípicas

A anotação dos aspectos nos sumários do CSTNews descrita em Rassi et al. (2013) foi inteiramente revisada neste trabalho. Cada um dos 50 sumários que compõem o *corpus* foi analisado manualmente após a revisão da literatura descrita na subseção anterior.

Apesar de Rassi et al. (2013) elencarem uma série de dúvidas a respeito da identificação dos aspectos, não foi possível propor nenhuma mudança na anotação do aspecto COUNTERMEASURES. Por exemplo, segundo Rassi et al. (2013), a anotação da sentença *"O governo de Seul manifestou sua predisposição em oferecer ajuda à Coréia do Norte"* com a etiqueta de macroaspecto COUNTERMEASURES é duvidosa, posto que essa sentença não parece veicular informação totalmente condizente com a definição dada ao aspecto em questão, que é: "medidas que visam solucionar/antecipar/impedir problemas relacionados a um fato".

No entanto, não foi possível estabelecer outro aspecto que mais adequadamente representasse o conteúdo da sentença. Isso talvez se deva ao fato de que esse caso é isolado; todos os outros casos de COUNTERMEASURES condizem com a definição proposta para o aspecto.

Além desse caso, outro que foi apontado como problemático por Rassi et al (2013), é o da anotação de COMMENT na sentença *"A seleção de vôlei de Bernardinho manteve sua hegemonia mundial derrotando a Rússia, mesmo perdendo o primeiro set".* Segundo os autores, essa sentença, originalmente anotada com o aspecto COMMENT por causa do trecho *"mesmo perdendo o primeiro set"*, poderia expressar, na verdade, um HOW, ou seja, "modo como um fato ocorreu".

Tendo em vista que COMMENT ocorre exclusivamente como macroaspecto na categoria "esporte", reconsiderou-se a anotação da sentença em questão com a etiqueta COMMENT por causa do trecho específico *"mesmo perdendo o primeiro set".* Assim, sugere-se, como resultado da análise, a troca da etiqueta COMMENT por HOW, particularmente porque o conteúdo é veiculado por um trecho sentencial específico.

Com exceção dessa sugestão de mudança na anotação, a revisão manual dos aspectos nos sumários confirmou o rigor teórico-metodológico de Rassi et al. (2013), culminando na manutenção do rol inicial de aspectos e rótulos e das organizações prototípicas propostas pelos autores quando da anotação do *corpus* CSTNews.

A seguir, descreve-se a investigação sobre as características linguísticas dos microaspectos que podem subsidiar a identificação automática dos mesmos juntamente com os atributos utilizados na anotação de papéis semânticos.

4. Caracterização dos microaspectos para detecção automática

Nesta subseção, apresenta-se a análise das ocorrências no CSTNews de alguns dos microaspectos.

Da análise dessas ocorrências, foram levantadas características linguísticas dos aspectos com potencial para subsidiar estratégias de detecção automática. Uma vez delimitadas, as características foram traduzidas em regras no formato lógico *se, então,* tratável por máquina.

Especificamente, essas regras codificam as condições (*se*) de ocorrência dos aspectos nos sumários e, por conseguinte, a anotação *(então*) a ser feita. Assim, dado um texto novo, uma ferramenta automática de anotação de aspectos sabe que, se uma das condições ocorrer, há um aspecto a anotar.

Em especial, do conjunto de 7 microaspectos (cf. Quadro 2 e 6), foram analisados os seguintes: (i) WHEN e WHY, comuns a todas as categorias do *corpus*, (ii) SCORE, específico da categoria "esporte" e (iii) SITUATION, que não possui um papel semântico correspondente.

4.1. Especificação de regras para a identificação de WHEN

No Quadro 1, WHEN é o microaspecto definido por Rassi et al (2013) aquele que expressa "a data/período de tempo (estritamente temporal) da ocorrência de um fato/evento".

Nos sumários do CSTNews, há 40 ocorrências de WHEN, quando ligado ao evento principal, o que equivale a 5% dos aspectos anotados. Na correlação com os papéis semânticos do PropBank-Br, WHEN é similar a ArgM-TMP (argumento temporal) (isto é, tempo em que a ação ocorreu). Ao analisar as ocorrências de WHEN, verificou-se que, sintaticamente, ele é um sintagma preposicional (SPrep), que pode ter diferentes configurações internas.

No Quadro 8, as configurações dos SPreps que expressam WHEN são formalizadas em regras lógicas (*se, então*) e exemplificadas com ocorrências do *corpus.*

Quanto ao Quadro 8, vale ressaltar que os símbolos "+", "+/-" e "|", significam, respectivamente: "seguido de", "seguido ou não de" e "ou".

Para a aplicação das regras do Quadro 8, no entanto, uma ferramenta automática de anotação de aspecto requer a consulta a um dicionário em que estejam armazenadas as informações necessárias para o reconhecimento na sentença dos elementos que compõem as regras.

Assim, além das preposições (p.ex.: *de, em, a,* etc.), artigos (*a(s), o(s), um, uns, uma, umas, à(s)*), pronomes (p.ex.: *ele(s), ela(s), este(s), esta(s), esse(s), essa(s), aquele(s), isto, isso,* etc.) e numerais (*1, 2, um, dois,* etc.), comumente presentes nos léxicos computacionais, a aplicação das regras do Quadro 8 requer a consulta a um dicionário ou léxico que também contenha as entradas propostas no Quadro 9. Tais entradas contêm conjuntos limitados de palavras.

Quadro 8: Regras formais para a detecção automática do microaspecto WHEN.

Regra	Condição (Se)/Ação (Então) + Exemplo
1	*Se* S tem [PREPOSIÇÃO + (PRONOME|ARTIGO) + ADVÉRBIO_DE_TEMPO + PREPOSIÇÃO + (PRONOME|ARTIGO) + DIA_DA_SEMANA + NUMERAL], *então* anotar S com WHEN P.ex.: A chuva complicava o trânsito [na_PREPOSIÇÃO+ARTIGO manhã_ADVÉRBIO_DE_TEMP0 desta_PREPOSIÇÃO+PRONOME segunda-feira_DIA_DA_SEMANA, 16_NUMERAL]
2	*Se* S tem [PREPOSIÇÃO + (PRONOME|ARTIGO) + DIA_DA_SEMANA], *então* anotar S com WHEN P.ex.: Um terremeto atingiu o Japão [nesta_PREPOSIÇÃO+PRONOME segunda-feira_DIA_DA_SEMANA, matando 9 pessoas.]
3	*Se* S tem [PREPOSIÇÃO + (PRONOME|ARTIGO) +/- (NUMERAL|TOKEN) + LÉXICO_DE_TEMPO], *então* anotar S com WHEN P.ex.:[Aos_PREPOSIÇÃO+ARTIGO 18_NUMERAL minutos_LÉXICO_DE_TEMPO], Maicon fez o primeiro gol. [No_PREPOSIÇÃO+ARTIGO primeiro_TOKEN tempo_LÉXICO_DE_TEMPO] houve outras jogadas [...]. Os acontecimentos ocorreram [nessa_PREPOSIÇÃO+PRONOME semana_LÉXICO_DE_TEMPO].
4	*Se* S tem [PREPOSIÇÃO + (PRONOME|ARTIGO) +/- (TOKEN|NUMERAL) + ADVÉRBIO_DE_TEMPO], *então* anotar S com WHEN P.ex.: A quarta medida foi aprovada [nesta_PREPOSIÇÃO+PRONOME madrugada_ADVÉRBIO_DE_TEMPO]

Quadro 9: Dicionário de suporte à aplicação das regras de detecção do microaspecto WHEN.

DIA_DA_SEMANA = [segunda-feira, terça-feira, quarta-feira, quinta-feira, sexta-feira, sábado, domingo]
ADV_TEMP = [hoje, amanhã, ontem, anteontem, tarde, madrugada, noite, meia-noite, manhã]
LÉXICO_DE_TEMPO = [microssegundo, segundo, minuto, hora, dia, semana, mês, ano, década, milênio, semestre, bimestre, trimestre, época, tempo, set]

4.2. Especificação de regras para a identificação de WHY

Conforme o Quadro 1, WHY é o microaspecto definido por Rassi et al (2013) aquele que expressa "a explicação de o porquê um fato/evento acontece (ou aconteceu)". Nos sumários do CSTNews, há 18 ocorrências de WHY, quando ligado ao evento principal, o que equivale a 2% do total de aspectos anotados. Na correlação com os papéis semânticos do PropBank-Br, WHY é similar a ArgM-PRP (propósito) (isto é, "motivação da ação expressa pelo verbo"). Ao analisar as 18 ocorrências de WHY, constatou-se que ele é frequentemente materializado por expressões específicas que indicam "causa". No Quadro 10, essas expressões de WHY são formalizadas em regras lógicas (*se, então*) e exemplificadas com ocorrências do *corpus.*

Quadro 10: Regras formais para a detecção automática do microaspecto WHY.

Regra	**Condição (*Se*), Ação (*Então*) + Exemplo**
1	*Se* S tem um elemento constitutivo do LÉXICO_DE CAUSA, *então* anotar S com WHY P.ex.: O presidente morreu [por causa disso_LÉXICO_DE_CAUSA]. [Com isso__LÉXICO_DE_CAUSA], o presidente morreu. O presidente sobreviveu [graças à_LÉXICO_DE_CAUSA] ajuda médica.
2	*Se* S tem a preposição "*por*" + VERBO_INFINITIVO, *então* anotar S com WHY. P.ex.: O presidente morreu [por_PREPOSIÇÃO beber_VERBO_INFINITIVO] vinho envenenado. [Por_PREPOSIÇÃO comer_VERBO_INFINITIVO] veneno, o presidente morreu.

Com base no Quadro 10, vê-se que, além do reconhecimento da preposição frequente em português "*por*" e das formas verbais no infinitivo (p.ex.: "*comer*", "rir", "amar", etc.), a aplicação das regras requer a consulta a um dicionário ou léxico que também contenha uma entrada composta por um conjunto limitado de expressões que indicam causa, como a ilustrada no Quadro 11.

Quadro 11: Dicionário de suporte à aplicação das regras de detecção do microaspecto WHY.

LÉXICO_DE_CAUSA =	[por isso, com isso, porque, devido a, por causa de, graça(s) a/ao, por força de, em função de, em virtude de, em razão de, em decorrência de, em consequência de, pois, visto que, já que]

4.3. Especificação de regras para a identificação de SCORE

Conforme o Quadro 1, SCORE é o microaspecto definido por Rassi et al (2013) como "resultado numérico de um fato/evento (*score*, tempo, distância, etc., sobretudo relativo a esportes)".

Nos sumários do CSTNews, há 7 ocorrências de SCORE, quando ligado ao evento principal, o que equivale a 1% do total de aspectos anotados. Na correlação com os papéis semânticos do PropBank-Br, SCORE é similar a ArgM-EXT (quantidade) (isto é, quantidade da mudança da ação). Ao analisar as 7 ocorrências de SCORE, constatou-se que é materializado na superfície do texto por expressões fixas, que expressam resultados de esportes específicos.

No Quadro 12, tem-se regras que buscam formalizar a estrutura dessas expressões, as quais são exemplificadas com ocorrências do *corpus.*

Quadro 12: Regras formais para a detecção automática do microaspecto SCORE.

Regra	Condição (*Se*), Ação (*Então*) + Exemplo
1	*Se* S tem [PREPOSIÇÃO + 3/2/1 + TOKEN("*sets*") + TOKEN("*a*") + 2/1/0], *então* anotar SCORE P.ex.: [...] venceu a Finlândia [por_PREPOSIÇÃO 3 sets_TOKEN a_TOKEN 0], em Tampere (FIN) [...] [...] derrotou a Rússia [por_PREPOSIÇÃO 3 sets a 1] [...]
2	*Se* S tem [PREPOSIÇAO + NUMERAL + TOKEN("*a*") + NUMERAL], *então* anotar SCORE P.ex.: [...] [por_PREPOSIÇÃO 3_NUMERAL a_TOKEN 0_NUMERAL]
3	*Se* S tem [NUMERAL + TOKEN("*m*"/ "*metro*"/ "*metros*") + NUMERAL] P.ex.: [...] conquistou a medalha de ouro no salto com vara ao saltar [4_NUMERAL m_TOKEN 60_NUMERAL] [...]
4	*Se* S tem [NUMERAL + TOKEN("*min*"/"*minuto*"/"*minutos*") + NUMERAL + TOKEN("*s*"/ "*segundo*"/ "*segundos*" + NUMERAL] P.ex.: [...] ao cravar o tempo de [7_NUMERAL min_TOKEN 12_NUMERAL s_TOKEN 27_NUMERAL [...] [...] com o tempo de [3_NUMERAL min_TOKEN 15_NUMERAL s_TOKEN 90_NUMERAL [...]

4.4. Especificação de regras para a identificação de SITUATION

Enquanto microaspecto, ressalta-se que SITUATION não possui um papel temático correspondente e, por conseguinte, não pode ser identificado por meio das estratégias ou atributos comumente utilizados na APS.

Dessa forma, fez-se uma análise das ocorrências desse aspecto no CSTNews com o objetivo de se identificar características linguísticas que pudessem subsidiar estratégias automáticas de identificação. De todas as 14 ocorrências de SITUATION (cf. Quadro 4), em 8 delas SITUATION se caracteriza como microaspecto. Especificamente, 2 casos de SITUATION ocorreram em sumários da categoria "mundo" e os outros 6, em sumários da categoria "esporte".

Do ponto de vista semântico, SITUATION em ambas as ocorrências de "mundo" referem-se a "ações limitares". Em uma delas, SITUATION é expresso pelo trecho "*nesta batalha*" e, em outra, pelo trecho "*conflitos entre o Exército e a guerrilha dos Tigres de Libertação da Pátria Tâmil*".

Do ponto de vista sintático, o trecho "*nesta batalha*" é um Sprep composto por [*em*_PREPOSIÇÃO + *esta*_PRONOME_ARTIGO + *batalha*_LÉXICO_MILITAR]. O outro trecho é um SN complexo, composto por um NOME (LÉXICO_MILITAR) + SPrep (= *conflito*_ LÉXICO_MILITAR + (*entre o Exército e a guerrilha dos Tigres de Libertação da Pátria Tâmil*)_SPREP].

Com base apenas nesses 2 casos do aspecto na categoria "mundo", pode-se dizer que SITUATION caracteriza-se em nível lexical, ou seja, pela ocorrência de palavras que expressam tipos de "ação militar", como os nomes "*batalha*" e "*conflito*".

No Quadro 13, tem-se regras que buscam formalizar as ocorrências de SITUATION na categoria "mundo".

Quadro 13: Regras formais para a detecção do microaspecto SITUATION na categoria "mundo".

Regra	**Condição (*Se*), Ação (*Então*) + Exemplo**
1	*Se* S tem [PREPOSIÇAO + (PRONOME|ARTIGO) + LÉXICO_MILITAR], *então* anotar SITUATION P.ex.:[Nesta_PREPOSIÇÃO+PRONOME batalha_LÉXICO_MILITAR], 15 soldados israelenses morreram [...]
2	*Se* S tem [LÉXICO_MILITAR + PREPOSIÇÃO], *então* anotar SITUATION P.ex.: [...] conflito_ LÉXICO_MILITAR entre_PREPOSIÇÃO o Exército e a guerrilha [...]

Com base no Quadro 13, vê-se que, além do reconhecimento das preposições, pronomes e artigos, a aplicação das regras requer a consulta a um dicionário ou léxico que também contenha uma entrada composta por um conjunto limitado de palavras que expressam ações militares, como a ilustrada no Quadro 14.

Quadro 14: Dicionário para aplicação das regras de detecção de SITUATION na categoria "mundo".

LÉXICO_MILITAR =	[ação, ação militar, batalha, combate, conflito, confronto, duelo, embate, guerrilha, luta, guerra]

A seguir, em (6), tem-se os 6 casos de SITUATON na categoria "esporte".

(6) a. [...] na Liga Mundial de Vôlei-06.] (C8)
b. [...] nos Jogos Pan-Americanos.] (C38)
c. [...] nos Jogos Pan-Americanos [...] (C41)
d. [...] em uma das provas mais charmosas da natação, o revezamento 4x100m livre [...] (C41)
e. [...] no salto com vara [...] (C24)
f. [Na estreia da seleção de vôlei do Brasil no Pan-Americano [...] (C48)

Nesses casos, o microaspecto SITUATION é expresso por um SPrep, composto por [Preposição + SN].

Nas ocorrências (6a), (6b) e (6c), os SPreps são introduzidos pela preposição "*em*" e os SNs caracterizam-se pela estrutura [artigo + nome próprio]. Nesses casos, os nomes próprios referem-se a eventos esportivos específicos ou gerais, a saber: "*Liga Mundial de Vôlei-06*" (6a) e "*Jogos Pan-Americanos*" (7b,c). Esses nomes próprios, em especial, são menções na superfície textual a entidades da categoria evento (esportivo). Em outras palavras, tais nomes, assim como qualquer outro nome próprio que indique local, pessoa, etc., podem ser denominados "entidades

nomeadas" (ENs). Em (6d) e (6e), os SPreps também são introduzidos pela preposição "*em*". No caso de (6d), o SN "*uma das provas mais charmosas da natação, o revezamento 4x100m livre*" introduzido pela preposição "*em*" é bastante complexo, tendo como núcleo "*prova*". Em (6e), o SN "*o salto com vara*" é mais simples, tendo como núcleo "salto (com vara)". Em (6f), o SPrep também é introduzido pela preposição "*em*" e o SN "*a estreia da seleção de vôlei do Brasil no Pan-Americano*", também é complexo, tendo como núcleo "*estreia*".

No Quadro 15, tem-se regras que buscam formalizar as ocorrências de SITUATION na categoria "esporte".

Quadro 15: Regras formais para a detecção do microaspecto SITUATION na categoria "esporte".

Regra	**Condição (*Se*), Ação (*Então*) + Exemplo**
1	Se *S* tem [PREPOSIÇÃO + (PRONOME|ARTIGO) + ENTIDADE_NOMEADA/evento_esporte], *então* anotar SITUATION P.ex.: [...] segunda medalha de ouro da natação brasileira [nos_PREPOSIÇÃO+ARTIGO Jogos Pan-Americanos_ENTIDADE_NOMEADA/evento_esporte] [...]
2	Se *S* tem [PREPOSIÇÃO + (PRONOME|ARTIGO) + (PREPOSIÇÃO|ARTIGO) + LÉXICO_ESPORTE], *então* anotar SITUATION P.ex.: [em_PREPOSIÇÃO uma_ARTIGO das_PREPOSIÇÃO+ARTIGO provas_LÉXICO_ESPORTE] mais charmosas da natação, o revezamento 4x100m livre.
3	Se *S* tem [PREPOSIÇÃO + (PRONOME|ARTIGO) + TOKEN("*estreia*")], *então* anotar SITUATION P.ex.: [Na_PREPOSIÇÃO+ARTIGO estreia_TOKEN da seleção de vôlei do Brasil no Pan-Americano, [...]

Com base no Quadro 15, vê-se que, além do reconhecimento das preposições, pronomes e artigos, a aplicação das regras requer a consulta a um dicionário ou léxico que também contenha entradas compostas como as ilustradas no Quadro 16. Ressalta-se que as entradas abaixo são de fato ilustrativas, pois, para que as regras do Quadro 15 sejam abrangentes, é preciso listar as principais entidades nomeadas da subcategoria "evento esportivo" e as palavras do campo semântico "esporte".

Quadro 16: Exemplo de léxico para aplicação das regras de detecção de SITUATION na categoria "esporte".

EN-evento_esporte =	[Liga Mundial de Vôlei, Jogos Pan-Americanos, etc.]
LÉXICO_ESPORTE =	[prova, estreia, salto com vara, etc.]

4. Considerações finais

Sobre as atividades de pesquisa realizadas, a revisão da literatura foi essencial para identificar correlações entre os aspectos informacionais e outros construtos linguísticos (a saber, papéis temáticos e movimentos ou papéis retóricos), o que permitiu compreender mais o objeto de estudo. Ademais, a análise das ocorrências de alguns microaspectos no CSTNews gerou estratégias/regras que têm o potencial de subsidiar o desenvolvimento de uma ferramenta de anotação de aspectos em textos em português. Por fim, ressalta-se que a revisão manual da anotação dos sumários do CSTNews confirmou a pertinência do rol de aspectos e a qualidade da anotação.

Referências

AFANTENOS, S.D., *et al.* Exploiting Cross-Document Relations for Multi-document Evolving Summarization. In: VOUROS, G. A., PANAYIOTOPOULOS, T. (Eds.). Methods and applications of Artificial Intelligence/ Helenic Conference on AI, 3, 2004, Samos, Greece. **Proceedings…** Samos, 2004. p. 410-419.

_____ *et al.* Using synchronic and diachronic relations for summarizing multiple documents describing evolving events. **Journal of Intelligent Information Systems**, Vol. 30, N. 3, pp. 183-226, 2008.

ALVA-MANCHEGO, F. **Anotação automática semissupervisionada de papéis semânticos para o português do Brasil**. São Carlos, 2013. 137p. Dissertação (Mestrado em Ciências da Computação) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2013.

BICK, E. **The parsing system PALAVRAS**: automatic grammatical analysis of portuguese in a constraint grammar framework. 2000. PhD Thesis. Arhus University, 2000.

BORBA, F.S. **Uma gramática de valências para o português**. São Paulo: Editora Ática, 1996.

CAMARGO, R. T. **Investigação de estratégias de sumarização humana multidocumento**. São Carlos, 2013. 117p. Dissertação (Mestrado em Linguística) – Departamento de Letras, Universidade Federal de São Carlos, São Carlos, 2013.

CARDOSO, P.C.F. et al. CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: RST BRAZILIAN MEETING, 3, 2011, Cuiabá/MT, Brazil. **Proceedings...** Cuiabá, 2011, p. 88-105.

DAYRELL, C., JR., A. C., LIMA, G., JR., D. M., COPESTAKE, A., FELTRIM, V., TAGNIN, S., E ALUISIO, S. Rhetorical move detection in English abstracts: Multi-label sentence classifiers and their annotated corpora. INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

(LREC´12), 8, 2012. Istanbul, Turkey. **Proceedings...** Istanbul, 2012, p. 1604-1609.

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

DURAN, M. S.; ALUÍSIO, S. M. Propbank-Br: a Brazilian Treebank annotated with semantic role labels. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC´2012), 8, 2012, Istanbul, Turkey. **Proceedings...** Istanbul, 2012, p. 1862-1867.

FILLMORE, C. J. The case for case. In: Bach, E., Harms, R. T. (Eds.). **Universals in linguistic theory**. Holt, Rinehart and Winston, Inc., p.1-88, 1968.

GARAY, A.Y.B. **Sumarização Multidocumento com base em aspectos**. São Carlos, 2013. 96p. Monografia de Qualificação (Mestrado em Ciências da Computação) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2013.

LAGE, N. **Estrutura da Notícia**. 5ª ed. São Paulo: Ática, 2002.

LI, P. *et al*. Generating Aspect-oriented Multi-Document Summarization with Event-aspect model. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2011, Edinburgh/Scotland. **Proceedings...** Edinburgh, 2011, p. 1137-1146.

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

MANN, W.C.; THOMPSON, S.A. **Rhetorical Structure Theory:** A Theory of Text Organization. Technical Report ISI/RS-87-190, 1987.

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

NEVES, M.H.M. **Gramática de usos do português**. São Paulo: Editora UNESP, 2000.

OWCZARZAK, K., DANG, H. T. Who wrote What Where: Analyzing the content of human and automatic summaries. In: ACL WORKSHOP ON AUTOMATIC SUMMARIZATION FOR DIFFERENT GENRES, MEDIA, AND LANGUAGES, Portland/USA, 2011. **Proceedings...** Portland, 2011, p. 25-32.

PALMER, M.; GILDEA, D.; KINGSBURY, P. The Proposition Bank: an annotated corpus of semantic roles. **Computational Linguistics**, 31 (1), p. 71-105, 2005.

RADEV, D.R. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In: ACL SIGDIAL Workshop on Discourse and Dialogue, 1, Hong Kong, 2000. **Proceedings...** Hong Kong, 2000, p. 74-83.

RASSI, A, P.; ZACARIAS, A.C.I.; MAZIERO, E.G.; SOUZA, J.W.C.; DIAS, M.S.; CASTRO JORGE, M.L.R.; CARDOSO, P.C.F.; BALAGE FILHO, P.P.; CAMARGO, R.T.; AGOSTINI, V.; DI-FELIPPO, A.; SENO, E.R.M.; RINO, L.H.M.; PARDO, T.A.S. Anotação de aspectos textuais em sumários do corpus CSTNews. **Série de Relatório Técnico do NILC**, NILC-TR-13-01. São Carlos-SP, Junho, 2013, 55p.

RAPOSO, E.P. **Teoria da gramática: a faculdade da linguagem**. Lisboa: Caminho, 1992.

STEINBERGER, J. *et al.* JRC's Participation in the Guided Summarization Task at TAC 2010. In: TEXT ANALYSIS CONFERENCE, 2010, Gaithersburg/Maryland/USA, 2010. **Proceedings...** Gaithersburg, 2010.

SWALES, J. M. (1990). **Genre Analysis: English in Academic and Research Settings**. Cambridge, UK: Cambridge University Press.

SWALES, J. M., FEAK, C. B. **Abstracts and the writing of abstracts**. Michigan: University of Michigan Press, 2009.

TEUFEL, S.; MOENS, M. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In: MANI, I., MAYBURY, M. T. (Eds.), **Advances in automatic text summarization**, MIT Press, 1999.

ZHOU, L.; TICREA, M.; HOVY, E. Multi-document Biography Summarization. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, Vancouver/Canada, 2005. **Proceedings...** Vancouver, 2005, p. 1-8.